González, Claudia M.

# La recuperación de información en el siglo XX : Revisión y aplicación de aspectos de la lingüística cuantitativa y la modelización matemática de la información

**Tesis presentada para la obtención del grado de Licenciada en Bibliotecología y Ciencias de la Información**

*Director: Prof. César O. Archuby*

UNIVERSIDAD NACIONAL DE LA PLATA

Facultad de Humanidades y Ciencias de la Educación

Departamento de Bibliotecología

# La recuperación de información en el siglo XX

## Revisión y aplicación de aspectos de la lingüística cuantitativa y la modelización matemática de la información

Tesina presentada para optar al grado de Licenciado por:

Bibl.Doc. **Claudia M. González**

Director: Prof. **César O. Archuby**

La Plata, Argentina, Noviembre 2007

# Tabla de contenidos

## Parte II: Técnicas de procesamiento textual



## Parte III: Trabajo experimental

# Introducción general

La Ciencia de la Información además de tener como objeto de estudio las propiedades, comportamiento y fuerzas que rigen el flujo de la información, se ocupa de perfeccionar los métodos de su procesamiento para una mejor accesibilidad y aprovechamiento. Dentro de ella, el área de la Recuperación de Información ha realizado un esfuerzo sostenido por más de cuatro décadas para desarrollar sistemas automáticos eficientes que permitan el tratamiento de una masa documental, que desde el medio impreso al digital, se ha proyectado en una multiplicidad de formas.

La preponderancia que el formato texto ha tenido hasta nuestros días, sumada a la imposibilidad de procesar con técnicas documentales manuales una cantidad cada vez más creciente de información, hizo que buena parte de las investigaciones y desarrollos en las Tecnologías de la Información se concentraran en el análisis del lenguaje natural para la búsqueda de soluciones. Puesto que el lenguaje y su uso constituyen una manifestación humana compleja, que puede abordarse desde los estudios morfológicos, sintácticos, gramaticales y discursivos; el verdadero problema para la búsqueda y recuperación de información es llegar a su dimensión semántica.

El conocimiento que se tiene o se debería tener de los métodos y las técnicas utilizadas en el corazón de estos sistemas, ha sido una preocupación en la formación de las nuevas generaciones de bibliotecarios del entorno digital. La complejidad que plantea la utilización de la matemática, la estadística y las probabilidades como único medio de modelizar problemas para hacerlos entendibles por las computadoras, ha dispuesto por largos años una barrera delante de esta disciplina de fuerte tradición humanística.

El objetivo principal de esta tesina es construir una masa crítica de conocimiento indagando en el ámbito de las Tecnologías de la Información sobre los diferentes desarrollos realizados en la interpretación automática de la semántica de textos y su relación con los Sistemas de Recuperación de Información. Partiendo de una revisión bibliográfica selectiva se busca sistematizar la documentación estableciendo de manera evolutiva los principales antecedentes y técnicas, sintetizando los conceptos fundamentales y resaltando los aspectos que justifican la elección de unos u otros procedimientos en la resolución de los problemas.

A fin de que todo lo expuesto no resulte muy distante del ámbito bibliotecario, se realizan algunos pequeños experimentos, que como material didáctico, pueden utilizarse para comprender elementos relevantes del tratamiento automático de la informacion.

El presente trabajo se estructura en tres partes. En la parte I, capítulos 1 y 2, se desarrolla brevemente las consideraciones generales sobre la Recuperación de Información y los Sistemas. Esto constituye el marco dentro del cual tienen lugar las técnicas estudiadas. También en esta parte, en el capítulo 3, se tratan los antecedentes más antiguos en lo que se refiere a estudios estadísticos del lenguaje y en el capítulo 4 algunos antecedentes que en los primeros tiempos de la computación, realizaron aportes significativos al tratamiento de los textos.

En la parte II se agrupan los capítulos que tienen que ver con el desarrollo de las técnicas seleccionadas. En el capítulo 5 se presentan las formas más básicas del tratamiento textual, desarrollando especialmente la temática de la reducción de las palabras a su raíz como forma de disminución de la variabilidad léxica de los textos. En el capítulo 6 se trata el tema de la ponderación de términos y ha sido seleccionado por ser recurrente en los diferentes modelos de Sistemas de Recuperación y porque además ha sido muy prolífico su tratamiento en la literatura de investigación. Por último, en el capítulo 7, se eligió exponer una técnica moderna, que valiéndose de los principios del algebra líneal y la operatoria de matrices, pretende obtener resultados de calidad semántica superior.

En la parte III se exponen los experimentos realizados con datos concretos. Se expone en cada caso el objetivo del experimento, la fuente de datos utilizada, los pasos seguidos en el procesamiento y los resultados obtenidos. Se ha elegido mostrar en el capítulo 8 la aplicación sobre datos propios, de procedimientos que confirmen lo sostenido por Zipf. Además se completa con una técnica de selección de términos particular. En el capítulo 9 se ha elegido mostrar el comportamiento del modelo vectorial de recuperación utilizando diferentes variantes en el tratamiento textual y aplicando diferentes funciones de similitud. En el capítulo 10 se presenta un breve estudio comparativo de los resultados obtenidos al aplicar un modelo basado en el Espacio Vectorial y un modelo basado en Semántica Latente.

# PARTE I

# CONTEXTO

# Y

# ANTECEDENTES

# Capítulo 1

# La automatización de la Recuperación de Información

## 1.1. Recuperación de Información (RI)

El término inglés "Information Retrieval" fue usado por primera vez por C. Mooers[1] en un escrito de 1952 y luego alcanzó popularidad con un trabajo de R. Fairthorne[2] de 1961 [1, p.2]. Se utilizó para designar un área emergente de investigación que después de la Segunda Guerra Mundial (1945) encontró dos razones suficientes para prosperar:

- En primer lugar, existía una necesidad creciente de resolver el problema del acceso físico e intelectual al conocimiento científico. Este conocimiento estaba registrado en una masa documental cuyo crecimiento exponencial se había dado en llamar "explosión documental"[3].
- En segundo lugar, surgía la tecnología computacional, que siendo capaz de procesar tanto números como textos, brindaba nuevas posibilidades al tratamiento de la información.

G. Salton [2, p.1] y R. Baeza-Yates [3, p.1] sostienen que la Recuperación de Información concierne a la representación, almacenamiento, organización y acceso a los items de información. P. Igwersen [4, p.49] manifiesta que tiene que ver con los procesos que involucran la representación, almacenamiento, búsqueda y hallazgo de información relevante a los requerimientos de un usuario humano. Si bien ambas definiciones proceden de paradigmas de investigación que abordan el problema de la RI desde diferentes perspectivas, clásico o algorítmico y cognitivista respectivamente [5, p.177-182], se observa que coinciden en los tres aspectos fundamentales: **representación**, **almacenamiento** y **acceso**.

---

[1] MOOERS, C.N. Information retrieval viewed as temporal signalling. En *Proceedings of the International Conference of Mathematicians*, Cambridge, August 30-September 6, 1950. Providence: American Mathematical Society, 1952, p. 572-573.

[2] FAIRTHORNE, R.A. *Towards information retrieval.* London: Butterworths, 1961.

[3] Esta situación se encuentra reflejada en el trabajo: BUSH, V. As we my think. *Atlantic Monthy,* 1945, 176, p.101-108.

La bibliografía especializada de origen anglosajón suele emplear diversas expresiones, a veces indistintamente, otras veces estableciendo claras diferencias: “Information retrieval”, “document retrieval”, “text retrieval”, “data retrieval”, “knowledge retrieval” o “factual retrieval”. Excede a los fines de este trabajo el establecer una precisa diferenciación entre ellas, pero si se puede sostener que:

- El término “information retrieval” es el de uso más general y suele involucrar a todas las demás designaciones.
- “Document retrieval” and “text retrieval” son términos conceptualmente cercanos, en el sentido de que los “textos” constituyen una de las grandes tipologías documentales. “Document retrieval” tiene una dimensión perceptiblemente más amplia ya que incorpora imágenes, audio, etc.
- Suele confrontarse “data retrieval” con “information retrieval” para diferenciar dos intenciones. Con la primera expresión se hace referencia a la recuperación de items almacenados en un sistema, dónde los datos tienen una estructura y semántica bien definida y son recuperados como respuesta a una interrogación puntual. Con la segunda se hace referencia más bien a la recuperación temática sobre textos en lenguaje natural, que no siempre son bien estructurados y pueden ser semánticamente ambiguos, donde la finalidad es recuperar información relevante para el usuario.
- “Fact retrieval” y “knowledge retrieval” son conceptos cercanos en el sentido que ambos se refieren a búsquedas no-documentales.

En el presente trabajo, el uso del término “Recuperación de Información” se ajusta a la dimensión documental -particularmente a la textual-, de acuerdo con la idea expuesta por F.Lancaster[4] (citado por C. van Rijsbergen [6, p.1]), quien sostiene:

> *«Information retrieval is the term conventionally, though somewat innacurately, applied to the type of activity discussed in this volume. An information retrieval system does not inform (i.e. change the knowledge of ) the user on the subject of his inquiry. It merely informs on the existence (or non-existence) and whereabouts of documents relating to his reques».*

Este marco de referencia es, sin duda, la visión más tradicional de la Recuperación de Información que se sintetiza en:

[4] LANCASTER, F. *Information retrieval systems: Characteristics, testing and evaluation.* Wiley: New York , 1968.

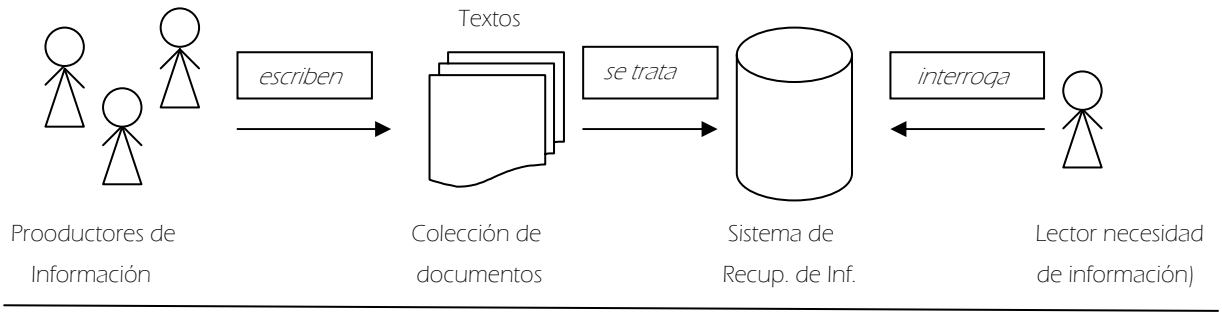



FIGURA 1: Modelo de RI

La creciente necesidad de información de las personas, derivada de algún tipo de requerimiento concreto: cultural, profesional, de investigación, educativo o recreativo, sumada a la constante producción de conocimiento plasmada por los autores en documentos; impone a los profesionales de la Bibliotecología y Ciencia de la Información[5] el principal desafío. Si de modo general se establece que el objeto de estudio de nuestra disciplina es el acceso a la información, por consiguiente, el fin último es lograr la comunicación efectiva entre los productores del conocimiento y los usuarios humanos. Uno de los medios tradicionales usados para facilitar esta comunicación son los Sistemas de Recuperación de Información (SRI), alrededor de los cuales se ha desarrollado una acumulación de saberes, abordados, frecuentemente, desde la perspectiva de diferentes disciplinas -psicología cognitiva, lingüística, computación- y que en su conjunto constituyen el área de estudio de la Recuperación de Información.

## 1.2. Los Sistemas de Recuperación de Información (SRI)[6]

Para C. van Rijsbergen [6¸ p.2] existen dos tipos de sistemas: los experimentales o de laboratorio y los operacionales o comerciales. La investigación en Recuperación de Información pretende entender el complejo proceso de la búsqueda de información con la finalidad de diseñar, construir y testar sistemas cada vez más eficientes. Sin embargo, es notable como, en toda la etapa previa a Internet, el desarrollo del área experimental ha sido mucho mayor que el nivel de transferencia logrado hacia sistemas concretos, de uso masivo. Para Blair [7, p.177] esto obedece a que, desde el origen mismo de las

[5] Usado en el sentido conceptual del término anglosajón “Library and Information Science (LIS)”.

[6] Dado que estos sistemas pueden ser manuales o automáticos y considerando que el lector conoce el nivel de desarrollo que los segundos han alcanzado, se aclara que en el presente trabajo cada vez que se nombre SRI se está haciendo referencia a aquellos que utilizan las computadoras para realizar algunos de sus procesos.

investigaciones, emergió el área de la **evaluación de la efectividad en la recuperación**[7] como metodológicamente muy consistente. Esto invitó a muchos a trabajar en el desarrollo de sistemas en laboratorio para demostrar mejores "performances" en los test de evaluación, más allá de un interés concreto en llegar a los usuarios[8].

### 1.2.1. El modelo general

Construir un modelo significa elaborar una representación simplificada de una realidad compleja con la finalidad de entenderla. Se utilizan para analizar, describir, explicar o exponer problemas, y su uso está muy difundido en la ciencia moderna [8].

La construcción de modelos en el área de la Recuperación de Información apareció a mediados de la década del setenta, con la aspiración de brindar una base teórica a la gran cantidad de experimentos y desarrollos que se estaban llevando acabo. Las numerosas técnicas diferentes y combinadas que empleaban, junto con la creciente necesidad de evaluar la efectividad de los sistemas, hacia imprescindible la construcción de patrones más generales en los cuales se pudiera agrupar la variedad de propuestas [1, p.257].

Un modelo sencillo de un proceso llevado a cabo en un SRI puede presentarse por la interacción de 3 elementos:

a- La necesidad de información de un usuario expresada en términos de una interrogación al sistema hecha en lenguaje libre o artificial[9].
b- La representación de la información contenida en el sistema realizada en términos de indización, categorias codificadas, representaciones gráficas.
c- La función de equiparación entre la interrogación y la representación de la información con la finalidad de encontrar coincidencias o similaridades.

Se puede visualizar en la siguiente figura los elementos que componen un SRI y sus interacciones [3, p.10]:

[7] Se refiere a los test de evaluación de sist. de indización y lenguajes de indización. Los primeros corresponden a la evaluación de la "performance" del sistema de unitérminos (ASTIA , USA y Cranfield, Reino Unido, 1953).

[8] La evaluación de los sistemas experimentales está ampliamente documentada en las Conferencias TREC [http://trec.nist.gov/]. La evaluación de un sistema operacional es de tipo "costo-baneficio".

[9] La distinción entre lenguaje libre y lenguaje artificial, así como otros conceptos relacionados con la indización, se abordan con mayor detalle en el apartado 1.3.1.

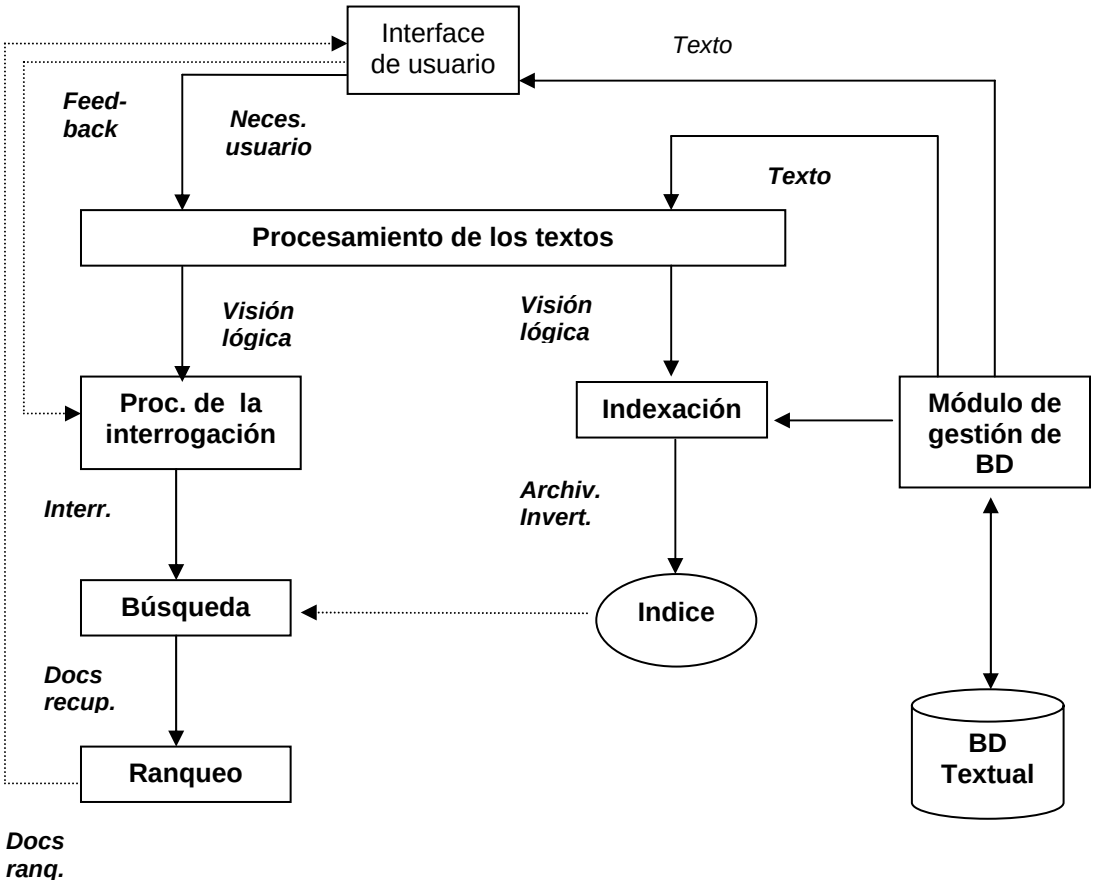



FIGURA 2: Modelo de SRI

Partiendo de un repositorio de información se construye una base de datos textual[10]. Para esto, se determinan los textos que se incluirán y las operaciones que se desarrollarán sobre ellos. Esto conformará la visión lógica del texto/documento. Luego que esto ha sido definido, y basándose en ello, el gestor de la base de datos construye el índice. Este será una de las partes centrales del sistema ya que es lo que permite acelerar los procesos de búsqueda de manera que sean viables en términos de volumen de información/unidad de tiempo. Una de las estructuras de índices más simple y muy usada es el archivo invertido[11]. Con estos elementos disponibles, el usuario puede iniciar un proceso de interacción con el sistema. Su necesidad de información es transformada por las mismas operaciones textuales aplicadas a los textos del repositorio y su necesidad es representada en términos

[10] Los SRI tradicionales están basados en una colección de items de información (documentos o partes de documentos) almacenados en una Base de Datos. Cada item constituye un registro de la base, el cual contendrá los valores de las características o atributos que los identifican. Estas bases se conocen, generalmente, como *Bases de Datos Bibliográficas* (hacen sólo referencia a los items existentes) y *Bases de Datos de Texto completo.* Estas bases de datos son creadas, mantenidas y controladas por los llamados *Sistemas de Gestión de Bases de Datos.*

[11] El archivo invertido contiene los términos de indización asignados. Cada término tiene asociado la lista de números de referencia a documentos que lo contienen. La recuperación de documentos identificados con algún término requiere que se busque primero en el archivo invertido el término y luego se llegue mediante la lista de referencias a los números de documentos en el archivo general.

de una interrogación concreta al sistema. Dicha interrogación permite, usando el índice ya generado, recuperar las representaciones de los documentos. Antes de enviárselas al usuario el sistema las ordena por probable relevancia[12]. A partir de allí, el propio usuario puede comenzar un ciclo de ajuste de su interrogación con la finalidad de recuperar los documentos más pertinentes para su requerimiento.

### 1.2.2. Modelos particulares

El desarrollo de modelos sobre un proceso tan complejo como el descrito anteriormente, en el cual interviene el hombre en toda su dimensión cognitiva y lingüística, además de las computadoras; puede hacerse desde diferentes perspectivas. Desde un punto de vista histórico, se puede decir que los primeros modelos se concentraron en describir los experimentos que se llevaban a cabo en relación con el mecanismo de recuperación del sistema, es decir la forma de equiparar la interrogación del usuario con el conjunto de documentos. Estos fueron los modelos **lógicos**, **vectorial** y **probabilístico** y son los que han apelado al lenguaje formal de la matemática como forma de describir con mayor detalle procesos realizables por las computadoras[13]. Estos tres modelos son los que más interesan en este trabajo, ya que es bajo su influencia que se han realizado los principales desarrollos en cuanto procesamiento del lenguaje natural.

En los últimos años se ha trabajado también en el desarrollo de modelos que enfocan la problemática del usuario y su interacción con la máquina. La forma en que la necesidad del usuario puede formularse como una interrogación que pueda ser procesada por un algoritmo. Estos modelos son los llamados **cognitivista** y **experto,** que conjuntamente con el **hipertextual** [1, p.257-258] son los que han reflejado las líneas de investigación en RI en la etapa inmediatamente anterior e inicial al desarrollo de la Web. En la actualidad, el

---

[12] Bookstein [9] define a la relevancia como la relación entre un individuo, en el momento que necesita una información, y el texto que se la provee. Se dice que el texto es relevante para esa persona si ella siente que la necesidad de información que tenía ha sido satisfecha al menos en parte por dicho texto. Sin embargo, esta noción de relevancia, intuitivamente correcta, no sería aplicable al entorno de los procesos automáticos desarrollados en un SRI dado que las decisiones son tomadas por una máquina, no por humanos. Se dirá entonces que existe otra noción para el concepto de relevancia que estaría signada por los procesos de carácter automático y que es el sentido con el que se la trata aquí. Dicha relevancia no es más que una estimación hecha por el SRI sobre el grado de adecuación del texto recuperado para con la interrogación del usuario. Los distintos modelos de SRI, en esencia, son diferentes formas de determinar la relevancia. T. Sarasevic [incluido en 1, p.143-165] utiliza dos conceptos diferentes para establecer la diferencia: **relevancia del usuario** y **relevancia del sistema**. J. Pérez Alvarez-Ossorio [10, p.63] propone tres conceptos: **relevancia formal**, que corresponde a la adecuación de la respuesta con la ecuación de búsqueda, la cual, salvo errores es siempre correcta; **relevancia semántica** que es la adecuación del pedido de información con lo obtenido, y por último, la **pertinencia** que tiene que ver con la adecuación a la necesidad del usuario.

[13] Robertson [11, p.128] sostiene al respecto «... works selected for discussion in this paper have a bias towards mathematical models, not because mathematics *per se* is necessarily a Good Thing, but because the setting up of a mathematical model generally presupposes a careful formal analysis of the problem and specification of the assumption and explicit formulation of the way in wich the model depends on the assumption».

paradigma parece ser la confluencia de lo mejor que ha dejado cada modelo más la incorporación de las “huellas” que dejan los usuarios al usar los sistemas.

## 1.3. Las técnicas automáticas de RI

Con la frase *técnicas automáticas* se hace referencia al conjunto de procedimientos y recursos que se aplican para que el sistema explote capacidades que el usuario no posee, lo alivie de las tareas rutinarias y trabajosas, o complemente y amplíe sus capacidades. Aplicando este criterio amplio, muchas de las cosas que se tratan en la bibliografía sobre RI son técnicas. Por ejemplo, la base de datos es un recurso de almacenamiento y recuperación de información que supera ampliamente a la memoria humana. Una interfaz de búsqueda gráfica es un procedimiento que permite el acceso temático a una colección de miles de documentos en solo 2 o 3 pantallas. Todas estas técnicas son empleadas para lograr una interacción exitosa entre un usuario, que puede ser humano o máquina, con cierta necesidad informativa; y una masa de información variada, registrada en documentos digitales o no, susceptible de satisfacer dicha necesidad y que puede o no haber sido sometida a algún proceso de descripción previo. Sin embargo, en este trabajo se acota el concepto de **técnicas automáticas de recuperación** a la clasificación propuesta por K. Sparck Jones [1, p.305], quien sostiene, no sin cierta dificultad, que las técnicas se pueden agrupar en:

- Técnicas de indización
- Técnicas de búsqueda

Las **técnicas de indización** tienen que ver con la construcción de la representación del documento y de la representación de la necesidad de información del usuario en el sistema de recuperación. Las **técnicas de búsqueda** tienen que ver con la manera en que el archivo de documentos es examinado y los ítems son extraídos de acuerdo a la interrogación que se formuló.

Una de las dificultades iniciales encontradas durante la elaboración de este trabajo obedeció a establecer con claridad la separación entre ambos tipos de técnicas, ya que se pretendía abordar en un primer momento exclusivamente las de indización. El solapamiento que muestra el tratamiento del tema en la bibliografía, además de encontrar desarrollos que optan por hacer la extracción de las palabras que representan al documento en el mismo momento de la búsqueda (en lugar de construir un índice que estuviera disponible “a-priori”), nos hace sostener que esta división es un tanto artificiosa. Es por ello que, frente a la

imposibilidad de tratar las técnicas de manera aislada, se ha optado por mantener el término tradicional de *técnicas de indización* y el interés en ellas en tanto tratamientos textuales contextualizados en los procesos de RI.

### 1.3.1. Las técnicas de indización

Para construir la representación de los documentos en el sistema de recuperación se puede optar entre una gran cantidad de metainformación[14]: autores, títulos, palabras del título, fecha de publicación, editores, fuente, tema sobre el que trata, partes del texto, etc. Generalmente se establece la diferencia entre la información bibliográfica/contextual y la información de contenido, aquella que dice sobre lo que el documento trata. Es en particular este tipo de información -la que está contenida en la semántica de los textos- la que interesa aquí, ya que marca la diferencia vital entre un Sistema de Recuperación de Información y un Sistema de Recuperación de Datos.

Para Blair [7, p.124] la significación de un texto reside en la relación que se establece entre el significante o expresión y el significado o contenido. La expresión es la portadora del contenido, y son las palabras o frases del lenguaje natural; mientras que el contenido es la imagen mental que se hacen los individuos al leer o escuchar una expresión determinada. La relación entre ambos, corresponde al concepto básico de signo lingüístico de Saussure[15]. Para la Bibliotecología y Ciencia de la Información, disciplina abocada a resolver problemas aplicados, el interés en el complejo problema de la significación radica en el uso que se hace de las expresiones para reflejar el significado de los documentos. La tarea de indicar sobre lo que trata el documento se conoce con el término **indización** y consiste en seleccionar algunos **términos[16]** para describir su contenido intelectual. Esta tarea puede realizarse a partir del propio vocabulario del documento, las mismas palabras empleadas por el autor; o bien, se puede utilizar un lenguaje documental o vocabulario de términos controlados o simplemente palabras elegidas por el indizador. El primer caso se denomina **indización derivada[17]** y el segundo **indización asignada**.

Cuando se utiliza el **lenguaje libre** del autor como forma de representación, se asume que existe un cierto nivel de conocimiento compartido entre productores de información y

[14] Metainformación: datos que decriben atributos de un recurso informativo. También se utiliza el término “Metadato”.

[15] SAUSSURE, F. *Curso de lingüística general.* Buenos Aires: Losada, 1968.

[16] Los términos son palabras o frases que sirven para designar exactamente un concepto determinado. El concepto es un elemento del pensamiento, conocimiento o juicio que refleja el resultado de la comprensión del mundo físico [13, p.54-56].

[17] En inglés se usa también el término “Natural Language Representation (NLR)”

usuarios. Igwersen [4, p.17] sostiene que cualquier procesamiento de información está mediado por un sistema de conceptos o categorías, un modelo del mundo que está generado y determinado por el conocimiento individual en un contexto social. Esto permite que la semántica del documento, lo que los sajones denominan "*aboutness*"[18] del autor, pueda ser representada con la misma terminología que el autor utiliza, ya que esta será, en cierta medida, reflejo del conocimiento colectivo. Cuando las interacciones de recuperación de información se realizan dentro de dominios temáticos específicos, se supone que existe mayor acuerdo terminológico [2, cap. 3], [12, cap. 3], [13, cap. 4].

Si la representación se construye utilizando un **lenguaje artificial** o documental, la indización es más selectiva y controlada. Al ajustar el vocabulario se busca solucionar los problemas de sinonimia y polisemia del lenguaje natural, así como las variaciones de las palabras.

La aplicación de un lenguaje documental elaborado “a-priori” o la selección de ciertos términos usados por el autor, podrá hacerse de diferentes maneras. La primera y más popular, es realizar la tarea intelectualmente para luego almacenar el resultado con apoyo informático. Esto se denomina **indización asistida por ordenador**. El paso siguiente, está dado por la extracción de los términos usados por el autor de manera automática, pero considerándolos solo como propuestos, ya que será un indizador humano quién tome la decisión final. Esto se denomina **indización semiautomática**. Por último, la **indización automática**, dónde la extracción se realiza automáticamente, pero también la validación de los términos como términos de indización definitivos [13, p.101]. Es posible también, aplicar lenguajes documentales automatizados: glosarios, tesauros, sistemas de clasificación. En este caso, el "*aboutness*" del autor es extraída automáticamente y conscientemente modificada de acuerdo a la estructura del vocabulario controlado. En este caso se dice que el conocimiento es tratado por dos agentes distintos: el autor y el diseñador del tesauro, es decir el experto en el dominio temático.

Blair sostiene que todas las formas potencialmente elegibles para construir la representación temática del documento procederán de la aplicación de alguno de estos métodos: **extracción, variación, descripción** [7, p.122]. En algún aspecto esto es coincidente con lo que manifiesta J. Moreiro González [13, p.17]

> *«Los distintos modelos de análisis documental automático han progresado desde una simple extracción de las palabras como objetos del lenguaje natural, pasando por un tratamiento lingüístico para la desambiguación de conceptos mediante la función cumplida por la palabra, hasta llegar a una indización “inteligente”, que trata de*

[18] Aboutness: sobre lo que trata el texto

*abstraer no solo conceptos sino modelos conceptuales fundamentados en bases de conocimiento, redes neuronales y sistemas avanzados de recuperación de información».*

En la Parte II de este trabajo se expondrán algunas técnicas de acuerdo con la clasificación propuesta por Blair:

- *Extracción.* Es necesario realizar sobre los textos ciertos tratamientos con la finalidad de reducir el vocabulario a tratar. Esto se hace extrayendo selectivamente (filtrando) el vocabulario más apropiado, el que sea más representativo de su semántica. Para lograrlo, algunas de las técnicas que se aplican tienen que ver con:
    1. Evaluar si todas las partes del texto son igualmente valiosas.
    2. Descartar caracteres o símbolos que no son de interés.
    3. Extraer las palabras que poseen carga conceptual de las que cumplen solo funciones de nexo.
    4. Aplicar métodos estadísticos y probabilísticos para:
        - Seleccionar las palabras que discriminan mejor al texto.
        - Valorar de manera diferenciada unas de otras (en un texto siempre hay un tema que se desarrolla más que otro).
    5. Procurar la extracción de términos compuestos y frases para lograr mejor calidad en la representación.

- *Variación*. Es necesario aplicar sobre el vocabulario extraído, cierta normalización en un nivel lingüístico, que permita avanzar hacia la desambiguación del lenguaje. Para lograrlo se aplican técnicas que tienen que ver con:
    1. Aplicar análisis léxicos que permitan la aplicación de algoritmos de reducción de las variantes de la palabra a una sola forma.
    2. Aplicar análisis sintácticos para reducir los problemas de homonimia según la ubicación de la palabra en la frase.

- *Descripción* Son técnicas que intentan lograr resolver problemas de nivel semántico superior

# Capítulo 2

# Los modelos de Sistemas de Recuperación de Información

Abordar en esta parte introductoria los diferentes modelos de SRI obedece a dos cuestiones principales. La primera es que se necesita exponer a nivel conceptual las ideas que han guiado a los experimentos de los cuales las técnicas de indización son parte. La segunda, de orden más práctico, es que ayuda a introducir al lector en los formalismos con que se expresan los procesos de RI para que su automatización sea posible.

La notación utilizada se basa en el modelo general propuesto por R. Baeza-Yates [3, p.23]. El sostiene que un modelo de SRI se define por cuatro elementos [D, Q, F, $R(q_i, d_j)$], dónde:

**1)** D es el conjunto de las representaciones de los documentos de la colección

**2)** Q es el conjunto de las representaciones de las necesidades de información de los usuarios

**3)** F es el marco dónde se modelan las representaciones de los documentos, las interrogaciones y las relaciones entre ambos

**4)** $R(q_i, d_j)$ es la función de ordenamiento por relevancia, la cual asocia un numero real con la interrogación $q_i \in Q$ y la representación del documento $d_j \in D$

para lograr mayor claridad en lo expuesto, se agregará que

- la interrogación $q_i$ se expresa con el conjunto de términos $k_a, k_b, k_c$ ....
- la representación del doc. $d_j$ se expresa con el conjunto de términos $t_a, t_b, t_c$ ...
- el conjunto de los documentos recuperados se denominará S

## 2.1. Modelos lógicos

### 2.1.1. Modelo Booleano

La idea principal de este modelo es que los documentos y las interrogaciones de los usuarios podrán ser representados por uno o más términos. Cada término debe pensarse como un **conjunto** cuyos elementos son los documentos que lo tienen asignado como término de indización. Se recordará que en el capítulo anterior se expuso que los sistemas

clásicos utilizan un índice de términos con la referencia a los documentos que lo poseen. Cuando se expresa el término $k_n$ en la interrogación, para el sistema se está expresando el conjunto de los documentos que lo contienen.

Si se pueden pensar los términos como conjuntos con elementos, se puede entonces establecer **operaciones** entre ellos. Los términos de la interrogación son susceptibles de ser enlazados por los **operadores** pertenecientes al **álgebra de Boole**:

- Y (AND) - Intersección
- (OR) - Suma o unión
- NO (NOT) – Resta o negación

El mecanismo de búsqueda del sistema es capaz de devolver todos los documentos que tienen la combinación de términos que satisfacen en términos lógicos a la interrogación. Así, para la colección de documentos

$$D=(d_1, d_2, d_3, d_4)$$

donde cada uno de los documentos tienen asignado los siguientes términos

$$d_1=(t_a, t_b, t_e, t_g, t_h)$$
$$d_2=(t_b, t_e)$$
$$d_3=(t_a, t_e, t_g)$$
$$d_4=(t_b, t_e, t_g, t_h)$$

las siguientes interrogaciones recuperarán los documentos del conjunto S

| | |
|---|---|
| $Q_1 = k_a$ AND $k_b$ AND $k_e$ | $S=(d_1)$ |
| $Q_2 = k_b$ AND $k_e$ | $S=(d_1, d_2, d_4)$ |
| $Q_3 = k_b$ OR $k_h$ | $S=(d_1, d_3, d_4)$ |
| $Q_4 = k_b$ NOT $k_h$ | $S=(d_2)$ |

considerando que los términos $k_n$ de la interrogación son equiparables a los términos $t_n$ de los documentos [2, p.25], [3, p.26], [6, p.74]

W. Cooper [14], enumeró algunos problemas que posee este modelo: la poca "amigabilidad" del lenguaje de interrogación[1]; la obtención de resultados nulos cuando se realiza una interrogación que incorpora muchos términos de búsqueda, o por el contrario, "sobre-recuperación" cuando se interroga al sistema utilizando muy pocos términos; la imposibilidad de que el usuario pueda destacar de alguna manera el término de búsqueda que más le interesa. Pero, sin duda, la principal crítica que se le realiza, es que en la función de equiparación subyace un criterio binario: el término de búsqueda está presente o está ausente en el documento, lo cual provoca por ejemplo, que en el caso de una operación de intersección (AND), el documento sea recuperado solo si todos los términos de búsqueda le fueron asignados. Si la interrogación posee cuatro términos de búsqueda, los documentos que solo poseen 3 o 2 de ellos no son recuperados. Por eso se lo suele llamar *modelo de equiparación exacta*.

Los sistemas que responden a este modelo suelen incorporar otras facilidades como el truncamiento, búsqueda por proximidad, búsqueda selectiva por campo o por rangos de fechas. Algunos sistemas que han empleado este modelo a lo largo de la historia de los SRI son: DIALOG, STAIRS, BRS, MEDLARS, ORBIT, LEXIS [2, p.30-45].

### 2.1.2. Mejoras al modelo Booleano

*Primeros pasos hacia la equiparación parcial*

La primer mejora que se incorporó a la función de equiparación en el marco de este modelo, es el concepto de **valor de punto de corte** [7, p.32] y consiste en que el usuario, además de brindar los términos de la interrogación, brinda un valor de las posibles combinaciones entre términos que también desea recuperar. Esto posibilita que la equiparación ya no sea rigurosamente exacta, lo que equivale a decir que algunos de los documentos pueden no tener algunos de los términos de búsqueda. Por ejemplo, para la interrogación:

$$q_1 = k_a \text{ AND } k_b \text{ AND } k_e$$

si se asigna un valor de corte=1, se recuperan no sólo los documentos que poseen los tres términos, sino también los que se correspondan con las siguientes combinaciones:

[1] Hace referencia por ejemplo, a la constante confusión entre los operadores Y (AND)/O(OR). El Y(AND), lingüisticamente transmite idea de agregación, cuando en realidad es un operador restrictivo.

$k_a$ AND $k_b$

$k_a$ AND $k_e$

$k_b$ AND $k_e$

La cantidad de combinaciones estará dada por la combinatoria de:

$$\sum_{i=r+1}^{n}$$

donde n=cantidad de términos de la interrogación y r=el valor de corte indicado.

El sistema realiza las combinaciones de términos automáticamente, es decir que el usuario se ve aliviado, en cierta medida, de realizar sucesivos intentos de búsqueda. Sin embargo, también presenta el problema de que usuarios inexpertos brinden muchos elementos de búsqueda en la interrogación y un valor bajo como valor de corte. Por ejemplo una búsqueda con 7 términos y un valor de corte de 4, produce 21 posibles combinaciones[2], lo que implica mayor esfuerzo de procesamiento. Además se recuperan muchas veces los mismos documentos ya que se equipararán con más de una combinación. Los sistemas se ven obligados a hacer esfuerzos extras para eliminar la redundancia.

Igualmente, se lo debe valorar como el primer paso para "suavizar" el comportamiento riguroso del sistema frente a la realidad del usuario, para quien los documentos no se pueden clasificar tan taxativamente en sirve/no sirve. Habrá algunos que sirvan más que otros.

### *Ordenamiento*[3]

La segunda mejora sustancial a este modelo proviene de la necesidad de lograr salidas ordenadas, de manera que se muestren en los primeros lugares aquellos documentos que cumplen mejor con la interrogación. Cómo ya se estableció en el apartado anterior, los sistemas comienzan a recuperar documentos que cumplen la condición de búsqueda plenamente y otros que la cumplen parcialmente. Los primeros pasos en este sentido, se dieron en ordenar la salida de acuerdo al grado de solapamiento entre los términos de búsqueda y los términos de los documentos, de manera que aparezcan primero los que poseen todos los términos de búsqueda, luego los que poseen todos menos uno, etc.

[2] Según lo expuesto, entonces, se deberá calcular la combinatoria de 7 términos de búsqueda tomados de a 5 (valor de corte + 1) $C_7^5 = \binom{7}{5} = \frac{7!}{5!(7-5)!} = \frac{7!}{5!*2!} = \frac{5040}{120*2} = 21$. Se recuerda que el factorial de un número n!=1*2*3*4* (n-1)*n.

[3] También denominado "ranking".

*Ponderación y ordenamiento*

La tercer mejora que queremos destacar aquí es quizá la más sustancial y tiene que ver con la posibilidad de que

- el usuario pueda destacar dentro de los términos de búsqueda aquellos que más le interesan.
- el indizador pueda destacar dentro de los términos que asigna al documento aquellos que son más importantes.

Esto es el origen de las llamadas funciones de ponderación que luego se aplicarán también al proceso de ordenamiento de la salida.

En el caso del usuario, este asigna un número positivo a cada uno de los términos de búsqueda, generalmente en una escala de 1 a 7, donde trata de reflejar el valor de importancia que para él poseen los diferentes términos en esa búsqueda. Este valor se denomina **Peso** y, por ejemplo, para la búsqueda

$$q_1 = k_k\,(7)\ \text{AND}\ k_j\,(4)\ \text{AND}\ k_m\,(2)$$

los documentos serán devueltos en orden decreciente de las sumas de los pesos

| | | |
|---|---|---|
| 4to | $d_1 = (k_a, k_b)$ | 0 |
| 2do | $d_2 = (k_a, k_k, k_o)$ | 7 |
| 1ro | $d_3 = (k_k, k_j, k_o)$ | 11 |
| 3ro | $d_4 = (k_b, k_j, k_m, k_o)$ | 6 |

Según Blair [7, p.39], la dificultad de este modelo radica en que para el usuario es difícil asignar los pesos cuando, por regla general, desconoce totalmente el funcionamiento interno y el efecto que tiene en la efectividad de la recuperación elegir un valor u otro.

Para el caso del indizador, el problema que se trata de resolver es la limitación de que, frente a un tema marginal, la decisión a tomar se limita a asignar o no el término. La posibilidad de ponderar le brinda mayor flexibilidad y seguridad al poder aplicar una escala de importancia entre los términos. Así, de manera similar al caso anterior, para la interrogación no ponderada

$$q_1 = k_k\ \text{AND}\ k_j\ \text{AND}\ k_m$$

los documentos serán devueltos en orden decreciente de las sumas de los pesos

| | | |
|---|---|---|
| 4to | $d_1 = (k_a(7), k_b(3))$ | 0 |
| 1ro | $d_2 = (k_a(4), k_k(6), k_o(1))$ | 6 |
| 2do | $d_3 = (k_a(1), k_k(4), k_o(6))$ | 4 |
| 3ro | $d_4 = (k_a(3), k_k(1), k_o(2))$ | 3 |

El uso conjunto de ponderación en la interrogación y ponderación en la indización, aplicando una misma escala de valores para los pesos, permite que el punto de vista del indizador y del usuario se encuentre representado en la ordenación de la salida. Para la interrogación

$$q_1 = k_k\,(7)\ \text{AND}\ k_j\,(2)\ \text{AND}\ k_m\,(4)$$

los documentos serán devueltos en orden decreciente de las sumas de los productos entre el peso del término en la búsqueda y el peso del mismo término en el documento

| | | |
|---|---|---|
| 4to | $d_1 = (k_a\,(7), k_b(3))$ | 0 |
| 1ro | $d_2 = (k_a\,(4), k_k(6), k_o(1))$ | 42 |
| 2do | $d_3 = (k_a(1), k_k(4), k_o(6))$ | 28 |
| 3ro | $d_4 = (k_a(3), k_k(1), k_j(2))$ | 11 |

Para Baeza-Yates [3, p.23] es tan importante la forma en que se realiza el ordenamiento de la salida que ha llegado a manifestar que “es lo que determina al modelo”.

Dado que existen diversos algoritmos de ponderación, y por lo antes expuesto, se considera que la ponderación forma parte de la construcción de la representación, estas técnicas se exponen con mayor detalle en la Parte II de este trabajo.

### 2.1.3. Modelo basado en la lógica difusa

La lógica Booleana es determinista: un elemento pertenece o no pertenece al conjunto. La teoría de conjuntos difusos parte de la idea de que los conjuntos no tienen límites bien definidos y que los elementos pueden presentar diferentes grados de pertenencia. Se propone asociar a cada elemento una función que puede tomar el valor del intervalo [0,1], dónde 0 indica que el elemento no pertenece a la clase y 1 que pertenece totalmente. Se considera que la pertenencia a una clase es gradual, no abrupta, en el sentido de “pertenece/no pertenece”.

Se define el conjunto difuso H como:

$$H=\{x \in U, f(x)/f(x)>0\}$$

dónde

x=elemento
U=conjunto
f(x)= función que establece el grado de pertenencia del elemento al conjunto

Las tres operaciones básicas que se establecen son el complemento, la unión y la intersección. De esta manera, cuando se plantea una intersección entre dos conjuntos difusos, un elemento pertenece a la intersección de ambos con el valor mínimo que poseía

ese elemento en cada uno de los conjuntos o con una función que es el producto de las funciones, u otra que se demuestre adecuada [15, p.221]. Si se plantea una unión, se considerará el valor máximo de la función. Si se plantea un complemento, será 1- el valor de la función:

$$A \text{ y } B=\{x, f_y(x)=\min(f_A(x), f_B(x))>0\}$$
$$A \text{ o } B=\{x, f_o(x)=\max(f_A(x), f_B(x))>0\}$$
$$\text{No } A=\{x, f_{no}(x)=(1-(f_A(x))>0\}$$

La lógica difusa aplicada a la Recuperación de Información se usa para indicar que el peso del término t en el documento $d_1$ representa el grado en el cual el documento es miembro del conjunto de documentos indizados con ese término. Dada una cantidad de conceptos $t_a$, $t_b$, $t_c$ ... que representan diferentes áreas temáticas, es posible entonces identificar cada documento brindando la función de pertenencia a esa clase conceptual:

$$d_1=(f_{ta}(d_1), f_{tb}(d_1), \ldots )$$

El peso 1, indica que es miembro pleno del conjunto de documentos, 0 que no es miembro del conjunto y los valores intermedios indican grados parciales de pertenencia. [2, p.421], [12, p.17], [5, p.152].

## 2.2. Modelo vectorial

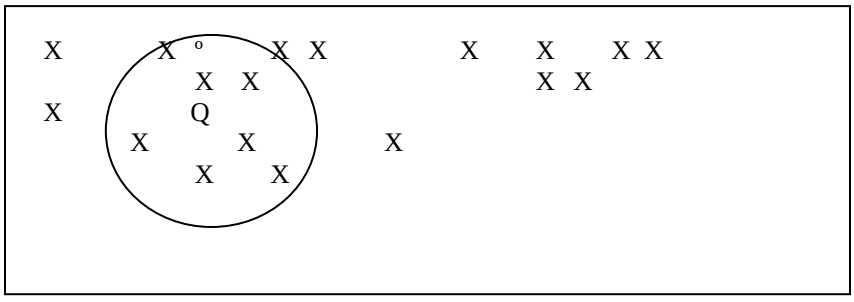
El principal cambio de concepción del modelo vectorial respecto al modelo Booleano, corresponde a la nueva visión que plantea respecto al espacio documental. Como ya se expuso, el modelo Booleano basa su mecanismo de recuperación en los que se denomina “índice invertido”. El espacio documental del sistema está dado por una lista de términos, ordenada de manera arbitraria, con sus correspondientes referencias al archivo de documentos general. El agrupamiento de documentos se da solo en el nivel de los términos individuales.

En el modelo vectorial el espacio documental se expresa de una manera diferente. Cada documento tiene uno o más términos asignados, pero la representación del documento en el sistema está dada por un **vector**, cuyos componentes serán los **pesos** que reflejen la importancia de cada término en ese documento en particular. Más allá de la manera más o menos sofisticada en que se calculen los pesos –que, como ya hemos manifestado se exponen en la Parte II de este trabajo -, al menos el peso tendrá valor 1 si el término está en

el documento o 0 si no lo está[4]. Lo importante es que cada documento está representado no solamente por los términos que contiene, sino también por los que no contiene. Se dice que es una representación n-dimensional, dónde n es la totalidad de los términos de indización del sistema. Así, se puede representar la colección de documentos mediante una **matriz término/documento**, dónde $P_{a1}$ es el peso del término $k_a$ en el documento $d_1$, $P_{a2}$ el peso del término $k_a$ en el documento $d_2$, etc.

| *Términos* | *Documentos* | | | | | | |
|---|---|---|---|---|---|---|---|
| | $P_{a1}$ | $P_{a2}$ | $P_{a3}$ | $P_{a4}$ | $P_{a5}$ | $P_{a6}$ | … |
| | $P_{b1}$ | $P_{b2}$ | $P_{b3}$ | $P_{b4}$ | $P_{b5}$ | $P_{b6}$ | … |
| | $P_{c1}$ | $P_{c2}$ | $P_{c3}$ | $P_{c4}$ | $P_{c5}$ | $P_{c6}$ | … |
| | … | … | … | … | … | … | $P_{ni}$ |

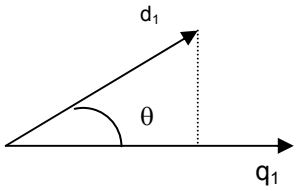
Si se piensa en cada documento como un punto en el espacio n-dimensional, dónde su posición está determinada por las coordenadas que indican los componentes del vector; realizando una abstracción, se puede convenir que existen conglomerados de puntos en aquellas zonas del espacio dónde existen documentos de temáticas más similares. Una gran parte del desarrollo de este modelo está dedicada a los cálculos de **similitud[5]** entre vectores. Esto se debe, en primer término, a que el modelo propone que la interrogación sea representada también como un vector de pesos susceptible de ubicarse en ese espacio n-dimensional. De esta manera, los puntos de los documentos que están más cerca del punto de la interrogación serán los más probablemente relevantes y deberán aparecer en los primeros lugares del orden de la salida. Se suele determinar un umbral que actuará en el sistema como valor del radio de una circunferencia que marca el límite de los documentos que se recuperarán:




[4] La asignación de pesos binarios, 0 o 1, no es lo más apropiado para este modelos. De hecho, las investigaciones que lo generaron siempre dedicaron especial esfuerzo al desarrollo de funciones de ponderación que son las que realmente l e brindan mejor "performance" en la recuperación.

[5] En la jerga profesional española suele usarse el término "similaridad", pero por ser un término no admitido en el RAE, se considera más apropiado utilizar similitud.

Una de las medidas de similitud que se encontró como satisfactoria, es un cálculo de la distancia entre los vectores midiendo su ángulo. Esta medida es llamada **coeficiente de correlación del coseno**[6] y muestra que cuando dos vectores son iguales, el ángulo entre ellos es de 0 grado, el coseno de 0 grado es 1. Si dos vectores son completamente diferentes, el ángulo entre ellos será de 90 grados (ortogonal). El coseno de un ángulo de 90 grados es 0. Entre 0 y 1 se darán todos los valores intermedios que reflejarán los diferentes grados de similitud entre las diferentes representaciones documentales y la representación de la interrogación.




$$sim(d_1, q_1) = \frac{\sum_{k=1}^{n} p_{(ka,d1)} * p_{(ka,q1)}}{\sqrt{\sum_{k=1}^{n} p^2_{(ka,d1)}} * \sqrt{\sum_{d=1}^{n} p^2_{(ka,q1)}}}$$

Esta misma medida, que aquí se presenta como una función de equiparación aplicable en el mecanismo de búsqueda, también es utilizada en el modelo vectorial para producir agrupamiento de documentos o **"clusters"**. Esto es un avance en el sentido de que ningún modelo de recuperación hasta ahora había superado la relación interrogación/documentos para incorporar en el modelo la relación inter-documentos. Esta técnica de agrupamiento es conocida como **clasificación automática**. Su utilidad está relacionada con una mayor eficiencia en el almacenamiento al guardar junto lo que tiene altas probabilidades de responder a la misma interrogación, es decir ser co-relevantes [1, p.259]. También ha sido empleada para producir interfaces gráficas de recuperación que permitan al usuario la búsqueda a partir de clases.

Todas las investigaciones llevadas a cabo por G. Salton alrededor del desarrollo del sistema SMART (1968) son las que dan origen a este modelo [7, p.44], [2, p.121, p.201], [3, p.27], [5, p.23], [15, p.247], [16], [17, p174].

[6] Otras formas se han empleado en los desarrollos de RI para cálculo de la similitud vectorial: Producto escalar, Coeficiente de Dice, Coeficiente de Jaccard, Coeficiente de solapamiento.

## 2.3. Modelo probabilístico

A mediados de los años 70 surgió una nueva manera de pensar el problema que los SRI debían solucionar. Los Investigadores W. Cooper, S. Robertson y K. Sparck Jones, retoman una idea que con anterioridad habían expuesto M. Maron y J. Kuhns en 1960 [18]. Se parte de suponer que la principal función de los sistemas de RI es ordenar los documentos de la colección en orden decreciente de probable relevancia ante la necesidad de información de un usuario[7]. Entonces, dada la consulta q se tratará de estimar la probabilidad de que el usuario considere al documento $d_1$ relevante, al documento $d_2$, al $d_3$, etc. Si el valor del cálculo de la probabilidad de $d_1$ es mayor que el de $d_2$, entonces $d_1$ será más relevante que $d_2$. El modelo asume que la relevancia de un determinado documento es independiente al resto de los documentos de la colección.

Se parte de la suposición de que existe dentro de la colección, un conjunto R de documentos relevantes ante la consulta q y un complemento de R de documentos no-relevantes ante la consulta q[8]. De la misma manera que en el modelo anterior, el documento está representado por un vector de términos con valores binarios: 1 si el término está presente, 0 si está ausente.

$$d_j = (t_a, t_b, t_c, \ldots t_n)$$

Entonces, se establece que la probabilidad[9] de que el documento $d_j$ sea relevante a la consulta q será:

$$P(R \mid d_j)$$

y que no lo sea

$$P(\overline{R} \mid d_j)$$

El problema es que no se conoce el valor del conjunto R inicial, que en principio el modelo lo asume como un conjunto ideal. Está claro que no se puede saber previamente si el documento en cuestión es relevante, entonces, habrá que encontrar una manera de

[7] W.Cooper define al Principio de Ordenamiento por Probabilidad como: «If a reference retrieval system´s response to each request is a ranking of the documents in the collections in order of decreasing probability of usefulness to the user who submitted the request, where the probabilities are estimated as accurately as possible on the basis of whatever data has been made available to the system for thid purpose, then the overall effectiveness of the system to its users will be the best that is obtainable on the basis of that data» [19].

[8] Se asume inicialmente la dicotomía Relevantes/No relevantes. Esto es visto como una de las desventajas del modelo [3, p.34].

[9] Ver nota al final del capítulo.

estimarlo. Esto, quizá, es el punto más crítico para la comprensión del modelo, pero debe asumirse así.

Ahora bien, se plantea la similitud entre $d_j$ y la interrogación q como la relación de la probabilidad de que sea relevante y que no lo sea.

$$sim(d_j, q) = \frac{P(R / d_j)}{P(\overline{R} / d_j)}$$

Esta misma función de similitud es expresada en términos de la Inversión de Bayes[10]

$$sim(d_j, q) = \frac{P(d_j \mid R) * (P(R)}{P(d_j \mid \overline{R}) * P(\overline{R})}$$

Aquí se interpreta que, dado que no se puede calcular la probabilidad de R condicionada a un $d_j$ en particular, se utiliza el teorema de Bayes para hacer la inversión, entonces se expresa la probabilidad ya no de R condicionada a $d_j$, sino la probabilidad de $d_j$ condicionada a R. Esto es: la probabilidad de un documento $d_j$ elegido aleatoriamente del conjunto R de los documentos relevantes. La probabilidad de R es la probabilidad de que cualquier documento sea relevante. Cómo la probabilidad de R y su complemento es igual en toda la colección (cualquier documento tiene la misma probabilidad de ser relevante o no ser relevante),

$$sim(d_j, q) \cong \frac{P(d_j \mid R)}{P(d_j \mid \overline{R})}$$

entonces se puede descartar ese valor ya que no nos agrega nada multiplicar numerador y denominador por una misma probabilidad:

Según van Rijsbergen [6, p.99], hay dos cuestiones que se deben asumir: la primera es que la probabilidad

$$P(R \mid d_j)$$

se puede lograr a partir de su inversión

$$P(d_j \mid R)$$

[10] Ver nota al final del capítulo.

La segunda, es que hay que considerar que la distribución de los términos de indización en los documentos relevantes es diferente que en los documentos no-relevantes, porque si se asumiera lo contrario, es decir que

$$P(d_j \mid R) = P(d_j \mid \overline{R})$$

no se podría establecer una función de discriminación entre los documentos de la colección que es una de las premisas en la Recuperación de Información.

Hasta aquí se ha planteado el tema de la probabilidad de relevancia de un documento de manera general. Asumiendo que los documentos son representados en este modelo por vectores de términos de indización independientes unos de otros[11], se plantea la similitud como:

$$sim(d_j, q) \cong \frac{(\prod_{g_i(d_j)=1} P(k_i \mid R)) * (\prod_{g_i(d_j)=0} P(\overline{k}_i \mid R))}{(\prod_{g_i(d_j)=1} P(k_i \mid \overline{R})) * (\prod_{g_i(d_j)=0} P(\overline{k}_i \mid \overline{R}))}$$

Finalmente, considerando que la probabilidad de que el término seleccionado esté presente en el subconjunto R, sumado a la probabiliad de que no lo esté, es igual a 1, es decir la probabilidad total:

$$p(k_i \mid R) + p(\overline{k}_i \mid R) = 1$$

Y aplicando logaritmos, se obtiene la función de equiparación utilizada en el modelo probabilístico:

$$sim(d_j, q) \cong \sum_{i=1}^{t} w_{i,q} * w_{i,j} * (\log \frac{P(k_i \mid R)}{1 - P(k_i \mid R)} + \log \frac{1 - P(k_i \mid \overline{R})}{P(k_i \mid \overline{R}}$$

Resumiendo, la función es una sumatoria de productos. Cada producto multiplica 3 cosas:

- el peso $w_i$ que tiene el término *i*-ésimo en el vector q de la interrogación
- el peso $w_i$ que tiene el término *i*-ésimo en el vector $d_j$ del documento
- una suma de las relaciones logarítmicas del *i*-ésimo término entre

[11] La independencia de los términos de indización es uno de los aspectos que se ha tratado de mejorar en este modelo. van Rijsbergen sostiene que es natural que los términos de indexación tegan que ver unos con otros y citando a Maron sostiene: «To do this [enlarge upon a request] one would need to program a computing machine to make a statistical analysis of index terms so that the machine will "know" wich terms are most closely associated with one another and can indicate the most probable direction in wich given request should be enlarged». Igualmente aquí se expone la versión más simple del modelo, dejando las funciones de pesado de los términos más sofisticadas para exponer en la Parte II de este trabajo.

- las probabilidades de que esté presente dentro de los documentos relevantes y que no lo esté
- las probabilidades de que esté ausente dentro de los no-relevantes y que no lo esté

Para poder aplicar esta función, el primer paso es encontrar una forma de calcular

$$P(k_i \mid R)$$

Una forma sencilla puede ser, en principio, asumir que es constante para todos los términos de indización y asignarle un valor de probabilidad, por ejemplo de 0,5. De forma similar, se puede asumir que la distribución de los términos de indización en el conjunto de los documentos no-relevantes puede aproximarse a la distribución de los términos en toda la colección de documentos

$$P(k_i \mid \overline{R}) = \frac{n_i}{N}$$

$n_i$ = Cantidad de documentos que poseen el término $k_i$

N= Cantidad de documentos en la colección

De esta manera, se hace la recuperación de los documentos que contienen los términos y se hace un primer ordenamiento. Luego se mejora recursivamente de la siguiente manera:

- El usuario establece un valor de corte en la salida de documentos que ya poseen un orden. Esto determina un conjunto de documentos V.
- Se delimita dentro de V el subconjunto $V_i$ que contienen el término $k_i$

$$P(k_i \mid R) = \frac{V_i}{V}$$

- Se aproxima el valor de la probabilidad de los relevantes a la distribución de $k_i$ dentro de los documentos recuperados
- Se aproxima la probabilidad de los no relevantes a que todos los no recuperados son no-relevantes

$$P(k_i \mid \overline{R}) = \frac{n_i - V_i}{N - V}$$

[9] Para lograr una mejor comprensión de lo que aquí se expone se cree conveniente en este punto repasar algunos conceptos básicos de probabilidades. Cómo se sabe, la probabilidad se aplica a cuestiones que implican un cierto grado de incertidumbre y consiste en obtener una estimación numérica de la posibilidad de que suceda o no suceda un determinado hecho. La definición general establece que la probabilidad de que ocurra un suceso determinado es igual al resultado de dividir el número de casos observados con determinada característica por la cantidad total de casos. Entonces, por ejemplo, i tomamos un dado, y decidimos calcular la probabilidad de que al tirarlo “salga un número par”, entonces se tiene que la probabilidad será el cociente 3/6. Dentro del universo

Ω=6 de números posibles, el evento A=3 son números pares. También la teoría de la probabilidad establece que, dados dos eventos A y B la probabilidad de que ocurran ambos, es la probabilidad del conjunto A intersección B:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) * P(B)}{P(B)}$$

Cuando A y B son eventos dependientes se habla de *Probabilidad Condicional*, es decir que la probabilidad de que se de A está condicionada a que B ocurra. En este caso será:

$$P(A) y P(B) = P(A \cap B) = P(A) * P(B)$$

En el caso de interés aquí, se establece que:

el universo Ω = colección de documentos
evento R = "que el documento $d_j$ sea relevante a la consulta q"
evento complemento de R = "que el documento $d_j$ no sea relevante a la consulta q"
evento $d_j$= "que el vector que representa al documento $d_j$ tenga algún tipo de similitud con el vector que representa la interrogación q"

Se observa que se expresa como una probabilidad condicional. La interpretación que aquí se dá, es que la probabilidad de que ocurra R (que sea relevante) está condicionada a que $d_j$ (cualquier vector que representa a un documento) tenga algo de similitud con q (la interrogación).
[20, p. 35]

[10] Partiendo de la probabilidad condicional, se puede decir que

$$P(B) * P(A \mid B) = P(A \cap B) = P(A) * p(B \mid A)$$

la intersección de A y B es lo mismo que expresar la intersección de B y A.

De aquí deriva la formula de la Inversión de Bayes:

$$P(A \mid B) = \frac{P(A) * (P(B \mid A)}{P(B)}$$

[19], [21, p.4].

# Capítulo 3

# Antecedentes del tratamiento automático de textos: era pre-computacional

En la primera parte de este trabajo ya se ha argumentado sobre la importancia del lenguaje en la Recuperación de Información. La automatización de cualquier proceso que lo involucre, implica que, necesariamente, se debe partir de la construcción de un modelo susceptible de ser expresado de manera lógico-matemática. Esto no es exclusivo del lenguaje, sino de cualquier problema que deba ser tratado por las computadoras, máquinas de base numérica por excelencia.

En el presente apartado se exponen, de manera selectiva, los antecedentes que se encuentran antes de la década de 1950 -momento del nacimiento de la informática- y que constituyen los primeros pasos hacia la visión del lenguaje como un fenómeno cuantitativo. La bibliografía cita a F.W. Kaeding[1] (1897), R.C. Eldridge (1911), J.B. Estoup (1916) y G. Dewey (1923), como los aportes más antiguos y a G.K. Zipf (1932)[2], (1935) [25], (1949)[3] cómo el que más trascendencia ha tenido, no solo por la continuidad que le dio al tema y la difusión con la publicación de sus libros, sino también porque llegó a la primera formalización matemática de algunos aspecto singulares del lenguaje. La obtención de bibliografía tan antigua no ha sido fácil, por lo cual se ha revisado aquí solo el trabajo de R.C. Elridge, J.B. Stoup, G. Dewey y parte del de G.K. Zipf.

## 3.1. R.C. Eldridge: la construcción de un lenguaje universal

La idea que movilizó a R.C. Eldridge en 1911 a realizar estudios de frecuencias de palabras fue la elaboración de un vocabulario reducido de uso universal [22]. Partió del concepto de que un número moderado de palabras, cuidadosamente seleccionadas, podía hacer que dos personas se entendieran aunque no tuvieran demasiado vocabulario en común. El primer paso que proponía para la construcción de dicho vocabulario consistía en seleccionar las palabras de los distintos idiomas que permitirían ese nivel de comunicación básico. El paso siguiente sería convertirlas en una única designación universal, para luego proceder a la eliminación de todas las otras palabras con significados similares en los distintos idiomas. El primer paso, el que corresponde a la selección de las palabras, es el que nos interesa comentar aquí.

[1] KAEDING, F.W. *Haufigkeitsworterbuch der deutschen Sprache.* Berlin: Ñ Selbstverlag des Herausgebers, 1897-98.
[2] ZIPF , G.K. *Selected studies of the principle of relative frequency in language.* Cambridge: MIT Press, 1932.
[3] ZIPF , G.K. *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley, 1949.

Si bien su propuesta era trabajar sobre diversos idiomas, su trabajo experimental fue realizado solamente en idioma inglés. El conjunto de datos expuestos muestra las palabras extractadas de 8 páginas de un periódico en distintas fechas, sobre diferentes temas y realizadas por diferentes escritores. Las 8 páginas contenían 250 artículos diferentes. Al analizar las tablas, Eldridge sostiene que dentro de las primeras 250 palabras de cada texto ordenadas por frecuencia, se encuentran prácticamente la totalidad de conectores del discurso. Si bien estas palabras son las más usadas, son para Eldridge las más difíciles de aprender cuando una persona se enfrenta a un nuevo idioma, dado que se usan de maneras muy diferentes. Sostiene que las primeras 750 palabras con sus repeticiones, constituyen más del 3/4 de las palabras de la lista y que esto debe darse de manera similar en cualquier texto. Entonces propone que en lugar de estas 750 palabras del inglés se construyera un vocabulario ecléctico con palabras de los idiomas más importantes, considerando que las palabras más adecuadas son las que poseen el mismo significado que las que se encuentran dentro de las primeras 750 del inglés. Luego, si este vocabulario ecléctico es injertado en los diferentes idiomas, y las palabras con significado similar son eliminadas, el mundo estaría frente a un idioma básico universal con la característica de equidad necesaria para que sea aprendido con los mismos niveles de dificultad por todos. Estas 750 palabras con sus repeticiones constituyen las 3 cuartas partes del vocabulario usado corrientemente. En promedio, entre 3 y 4 palabras de una oración en cualquier idioma se encuentran dentro de las 750 del vocabulario ecléctico.

## 3.2. J.B. Estoup: los estudios taquigráficos

En 1916, J.B. Estoup [23] propuso un método para el aprendizaje de la Taquigrafía que denominó “Gamas Taquigráficas”. Sostenía que el taquígrafo debía conocer uno a uno todos los signos de las palabras que han de estenografiarse. Para no desanimar a sus alumnos, realizó un pequeño estudio cuantitativo con la finalidad de demostrar que en realidad no son tantas las palabras que se usan en el discurso de un orador que improvisa, que es cuando adquiere sentido la taquigrafía.

En su estudio, conformó un texto con trozos de discursos de diferentes oradores con una longitud total de 30.000 palabras. Se cortó el texto en partes de 1000 palabras cada uno y se contó cuantas palabras diferentes tenía el primero, cuantas diferentes en el segundo que no estuvieran contenidas en el primero y así sucesivamente. Así se llega a la siguiente tabla:

| Trozos | Nros.palabras diferentes y nuevas | Por 100 | Término medio de Cuatro trozos |
|---|---|---|---|
| 1ro | 386 | 38,6 | |
| 2do | 253 | 25,3 | |
| 3ro | 210 | 21 | |
| 4to | **198** | 19,8 | |
| 5to | **132** | 13,2 | |
| 6to | **108** | 10,8 | |
| | | | |
| 7mo | **101** | 10,1 | **** |
| 8vo | **90** | 9 | **9,60** |
| 9no | **101** | 10,1 | |
| 10 | **92** | 9,2 | **/** |
| | | | |
| 11 | **99** | 9,9 | **** |
| 12 | **87** | **8,7** | **8,45** |
| 13 | **78** | **7,8** | |
| 14 | **71** | **7.4** | **/** |
| | | | |
| 15 | **76** | **7,6** | **** |
| 16 | **87** | **8,7** | **7,60** |
| 17 | **64** | **6,4** | |
| 18 | **77** | **7,7** | **/** |
| | | | |
| 19 | **86** | **8,6** | **** |
| 20 | **85** | **8,5** | **7,40** |
| 21 | **75** | **7,5** | |
| 22 | **50** | **5** | **/** |
| | | | |
| 23 | **64** | **6,4** | **** |
| 24 | **53** | **5,3** | **5,55** |
| 25 | **45** | **4,5** | |
| 26 | **60** | **6** | **/** |
| | | | |
| 27 | **43** | **4,3** | **** |
| 28 | **34** | **3,4** | **3,90** |
| 29 | **47** | **4,7** | |
| 30 | **32** | **3,2** | **/** |
| TOTAL | **2.987** | | |

TABLA I: Estudio de ESTOUP sobre uso de las palabras diferentes por los oradores.

Para formar el texto entero de 30.000 palabras bastaron 2.987. Afirma que si se siguiera con la tabla, probablemente al llegar al trozo de la posición 50 o 60, la proporción descendería a 1% o menos. Es decir que se llega a un punto en que se agota el vocabulario. A partir de las 30.000 palabras el aumento cada 10.000 se daría de la siguiente manera:

| | |
|---|---|
| De 31.000 a 40.000 | 3%, o sea 300 palabras |
| De 41.000 a 50.000 | 2%, o sea 200 palabras |
| De 51.000 a 60.000 | 1,5%, o sea 150 palabras |
| TOTAL | 600 palabras |

Es decir que un texto de 30.000 palabras más, solo sumaria 650 vocablos nuevos, lo que suma un total de3.637 para un texto de 60.000.

## 3.3. G. Dewey: la enseñanza del idioma inglés

En 1923, G. Dewey, profesor de la Universidad de Harvard, realizó un estudio sobre los sonidos simples y combinados más usados en el idioma inglés [24]. Para realizar el análisis fonético trabajó sobre un corpus de 100.000 palabras extraídas de materiales representativos del uso de un "buen inglés actual". Los materiales elegidos cubrían un gran espectro: editoriales y noticias de diarios, cuentos y novelas actuales, discursos, correspondencia personal y de negocios, avisos publicitarios, escritos religiosos y científicos, etc.

Los pasos que siguió la investigación fueron los siguientes: 1) contar, almacenar, ordenar y listar las 100.000 palabras, 2) Transcribir las diferentes palabras en notación fonética basada en el Alfabeto Científico Revisado, 3) Analizar las transcripciones fonéticas en relación con las sílabas. Reducir las representaciones a un alfabeto de 41 sonidos; ordenar, resumir y mostrar los datos de diversas formas, 4) Hacer lo mismo para los sonidos simples aislados.

Dewey consideró cada variación de las palabras con la misma raíz como palabras diferentes. Sin embargo, también realizó agrupamientos por variantes de una misma raíz. Se descartaron todo los nombres propios, números, títulos, abreviaciones, palabras extranjeras y signos de puntuación. Su estudio arroja datos como que:

| | | | | |
|---|---|---|---|---|
| 9 palabras | constituyen el | 25 % | de las 100.000 | palabras |
| 69 | " | 50 % | " | |
| 732 | " | 75 % | " | |

y en su libro registra más de 160 páginas con tablas de palabras y frecuencias.

Dewey considera que sus listas de palabras son de gran ayuda para la pedagogía de la escritura, así como el análisis de las sílabas y los sonidos ayudan a la evaluación de la lectura y a la corrección del deletreo.

## 3.4. G.K. Zipf: La distribución rango/frecuencia del lenguaje

Entre las décadas de 1930 y 1940, George Kingsley Zipf (1902-1950), también profesor de filología de la Universidad de Harvard, produjo una serie de trabajos que fueron tomados como pioneros por la lingüística cuantitativa y computacional para el estudio estadístico del lenguaje natural y su aplicación al desarrollo de algoritmos. Sus primeros trabajos están dedicados a demostrar ciertas regularidades del lenguaje, considerando a las palabras, fonemas y oraciones como simples eventos de la naturaleza susceptibles de ser analizados estadísticamente. Fue uno de los primeros en considerar al lenguaje como un proceso

biológico, psicológico y social, sobre el que cabía aplicar métodos científicos tradicionalmente usados en las ciencias exactas.

En un primer momento, Zipf, preocupado por los cambios fonéticos del lenguaje, se interesó en la frecuencia de uso de los fonemas en periodos largos de tiempo, para luego derivar su interés en la frecuencia de uso de las palabras. En 1932 publica su primer libro: "*Selected Studies of the Principle of Relative Frequency in Language"*, dónde mostró empíricamente cierta regularidad estadística observada en los textos a la que llamó "Principle of Relative Frequency" y que sería la base de sus trabajos posteriores. Según manifiesta Ronald Wyllys [25], dicho libro es un conjunto de diagramas y listados de palabras con sus frecuencias, en el cual, solo 22 páginas de las 125 páginas totales, están dedicadas a justificar sus observaciones.

En 1935 publica su segundo libro: *"The Psycho-Biology of Language: An Introduction to Dynamic Philology"* [26], donde profundiza sus estudios anteriores y establece gráficamente por primera vez lo que luego se conoce como la "Ley de Zipf". Su afirmación inicial es que:

> *existe una correlación estrecha entre la frecuencia con que ocurre una palabra y el número de las diferentes palabras que ocurren con esa frecuencia*

Esto quiere decir que si un número de palabras diferentes, considerando todas las inflexiones (no como unidades léxicas), ocurre una sola vez en el texto y lo designamos como X, el número de palabras diferentes que ocurren 2 veces, 3 veces, 4 veces, n veces es respectivamente $1/2^2$, $1/3^2$, $1/4^2$, ... $1/n^2$ de X [26, p.xii]. El valor de esta aseveración es que esto es válido en más del 95% de los casos. Así lo demuestra Zipf trabajando sobre 3 textos en diferentes idiomas: chino (un texto con 20 ejemplos de conversaciones en dialecto Peiping); en latín (4 diálogos de Plautus: Aulularia, Mostellaria, Pseudolus y Trinnumus) y en inglés (4 ejemplos de diarios Americanos compilados por R.C. Eldridge en una investigación previa [22]). Tal como se observa en las TABLAS II, III y IV al final de este capítulo, unas pocas palabras ocurren muchas veces, mientras que muchas ocurren raramente. Cuando el número de ocurrencias se incrementa, el número de palabras diferentes que poseen ese número de frecuencia decrece. Observando la TABLA IV para el idioma inglés, tomando solo las dos primeras columnas, la anterior aseveración queda expuesta en la TABLA V. Si se comparan los cambios que se van produciendo en los valores de la primera columna con los valores de la última, se observa, en este ejemplo, que son solo aproximados. Sin embargo, lo interesante aquí es que Zipf encontró que esta relación se mantiene bajo situaciones diversas, diferentes temas, autores e incluso idiomas. Zipf graficó de manera logarítmica este fenómeno para las 3 muestras anteriores (GRAFICO

II, III y IV al final del capítulo), tomando solo los valores de frecuencia de 1 a 45, es decir las frecuencias bajas. Así, observando el GRAFICO IV y su correspondiente TABLA IV, para el idioma inglés se ve que 2976 palabras (abcisa) ocurren 1 vez (ordenada), 1079 palabras ocurren 2 veces, etc. Ambos ejes están representados en escala logarítmica debido a que el rango de datos es muy amplio y este tipo de escalas permite observarlos en dimensiones y con detalles razonables. Zipf sostiene que la línea que se puede trazar entre los puntos marcados, responde a la fórmula $n * f^2 = c$ donde ***n*** representa el número de palabras de una frecuencia dada, ***f*** representa la frecuencia y ***c*** es una constante. Es decir que el producto del número de palabras de una frecuencia dada y el cuadrado de dicha frecuencia es aproximadamente constante para la mayoría de las palabras diferentes de un vocabulario en uso. Esto no se cumple para las palabras muy frecuentes. Si se observan los ejemplos de las 3 tablas anteriores, al final se encuentran agrupadas las altas frecuencias. Así, para la muestra en idioma chino, existen solo 12 palabras que varían entre las frecuencias 102/905; para la muestra en latín hay 71 palabras que varían sus frecuencias entre 62/514; y para el inglés 71 palabras que varían entre 61/4290. Para las muestras anteriores se puede observar el cálculo de la constante en la TABLA VI al final del capítulo. La relación $n * f^2 = c$ es válida solo para las palabras con frecuencias bajas, que por otro lado representan la mayor parte del vocabulario en uso. Es de este análisis inicial de donde deriva lo que comúnmente se conoce como “Ley de Zipf”. En *Psychobiology of language* Zipf manifiesta [2, p.44]*,*

> *«There is, however, another method of viewing and ploting these frequency distribution wich is less dependent upon the size of the bulk and wich reveals an additional feature. As suggested by a friend, one can consider the words of a vocabulary as ranked in the order of their frequency, e.g. the first most frequent word, the second most frequent, the five-hundredth most frequent, the thousandth most frequent, etc»*

De esta manera, que se podría pensar como casual, es como Zipf plantea una nueva manera de ver la distribución de uso de las palabras. Ya no dispone la relación entre la frecuencia y la cantidad de palabras con dicha frecuencia, sino una distribución de rango/frecuencia que es más general[4].

[4] Se pueden encontrar dos tipos de distribuciones que ordenan o “ranquean” un conjunto de entidades. La primera es la “distribución de rango”. Un ejemplo es cuando se le pide a un usuario que ordene el resultado de una búsqueda según sus preferencias. El problema que presenta es que el ordenamiento se hace en base al juicio, utilizando los números ordinales (primero, segundo, etc.) y la estadística trabaja con números cardinales. Entonces, a los efectos de la distribución se cambian unos por otros. El segundo tipo de distribución, es la “distribución rango/frecuencia” . En este tipo de distribución, las entidades exhiben cierta caracteristica que puede contarse en items o eventos. La frecuencia o ocurrencia de los items o eventos es determinada, y se ordena en forma decreciente. De forma contraria a la distribucion de rango, donde los ordinales se convierten en cardinales, en esta distribución, la frecuencia dada en números cardinales es la que ordena. Esto la hace más objetiva. La idea importante aquí es que las distribuciones de rango se basan en el juicio y las de “rango/frecuencia” en el valor de la frecuencia, y que ambas son de diferente tipo. Por ejemplo las

Así, sus observaciones anteriores, quedan expuestas:

> *Si se toma un texto y se cuenta la cantidad de veces que aparece cada una de las palabras (su **FRECUENCIA** de uso), luego se las ordena considerando dicho valor de forma descendente, y se le asigna **RANGO** 1 a la más frecuente, rango 2 a la segunda más frecuente y así hasta terminar; si después se multiplica el rango por la frecuencia, se observará que el resultado es aproximadamente **CONSTANTE*** [2, p.60] [26] [28] [29] [30, p.293] [31][32][33][34]
>
> $$r * f( r ) = c$$

Esta es la forma en que se difundió en el ámbito científico y es como la toman los autores del área de la Ciencia de la Información. Cuenta de ello lo dan los siguientes ejemplos extraídos de diferentes fuentes, que sirven para demostrar que otros autores han realizado experimentos similares. Si se analizan las siguientes tablas, se ve que el producto *r * f(r)* en realidad no es "exactamente" constante, sino una aproximación, mayor aún cuando la cantidad de datos analizados es más extensa. Si se observa la última columna en la TABLA VII, se determina que los valores son más constantes que en la TABLA VIII donde el corpus textual analizado es más pequeño

| Rango *(r)* | Palabra | Frecuencia *f (r)* | *r * (f(r)* /1.000.000) |
|---|---|---|---|
| 1 | the | 69.971 | 0.070 |
| 2 | of | 36.411 | 0.073 |
| 3 | and | 28.852 | 0.086 |
| 4 | to | 26.146 | 0.104 |
| 5 | a | 23.237 | 0.116 |
| 6 | in | 21.341 | 0.128 |
| 7 | that | 10.595 | 0.074 |
| 8 | is | 10.099 | 0.081 |
| 9 | was | 9.816 | 0.088 |
| 10 | he | 9.543 | 0.095 |

TABLA VII: Adaptación realizada por Salton sobre un corpus textual de 1.000.000 de palabras [2, p. 61].

| Rango *(r)* | Palabra | Frecuencia *f(r)* | *r * f(r)* |
|---|---|---|---|
| 1 | the | 245 | 245 |
| 2 | of | 136 | 272 |
| 3 | terms | 98 | 294 |
| 4 | to | 81 | 324 |
| 5 | a | 65 | 325 |
| 6 | and | 61 | 366 |
| 7 | in | 55 | 385 |
| 8 | we | 52 | 416 |
| 9 | request | 49 | 441 |
| 10 | documents | 40 | 400 |
| 20 | wich | 26 | 520 |

TABLA VIII: Extraída del artículo de A. Booth [35].

comparaciones que se pueden realizar en una distribución de rangos, no se pueden hacer en distribuciones de rango/frecuencia, mientras que los cálculos, solo son viables con números, no con ordinales. Se ha observado que en las actividades sociales que involucran un conjunto discreto y homogéneo de elementos -para la bibliotecología, por ejemplo, los libros o los lectores- las distribuciones de rango/frecuencia son logaritmicas. [27]

Para que el valor resultante fuera exactamente constante se debería dar que el valor de frecuencia del rango 2 fuera 1/2 del valor del rango 1, el valor de frecuencia del rango 3 fuera 1/3 del valor del rango 1, etc. Para los dos ejemplos anteriores, se puede observar en la TABLA IX la comparación entre valores reales e hipotéticos:

| *TABLA II* | | *TABLA III* | |
|---|---|---|---|
| *Frecuencia* | *valor hipotético* | *Frecuencia* | *valor hipotético* |
| 69.971 | | 145 | |
| 36.411 | *34,985* | 136 | *72,5* |
| 28.852 | *23,323* | 98 | *48,3* |
| 26.146 | *17,492* | 81 | *36,2* |
| 23.237 | *13,994* | 65 | *29* |
| 21.341 | *11,661* | 61 | *24,1* |
| 10.595 | *9,995* | 55 | *20,7* |
| 10.099 | *8,746* | 52 | *18,1* |
| 9.816 | *7,774* | 49 | *16,1* |
| 9.543 | *6,997* | 40 | *14,5* |

TABLA IX: Tabla de comparación de valores reales e hipotéticos de la TABLA II y III

Esta regularidad se observa mejor cuanto más extenso es el texto. Además, en tablas completas, que reflejan la totalidad de las palabras usadas en un texto, se aprecia claramente que existen unas pocas palabras que se repiten mucho, y una gran variedad que aparecen solo una vez

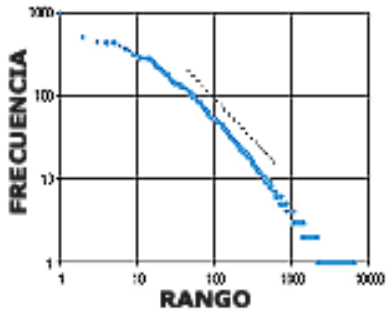



GRAFICO I: Curva de Zipf

Así mismo, Zipf diversifica la aplicación del conteo de frecuencias tratando de demostrar otro tipo de regularidades en los textos. Por ejemplo, establece que la distribución por frecuencias de las palabras del inglés demuestra que, la más frecuente ocurrirá un promedio de 1 cada 10 palabras, la segunda palabra más frecuente ocurrirá cada 20 palabras, la centésima palabra más frecuente ocurrirá cada 1000 palabras, la enésima palabra más frecuente, ocurrirá cada 10 * n palabras. De la misma manera, realizó estudios sobre la longitud de las palabras, llegando a establecer que las palabras cortas son más favorecidas en el uso que las palabras largas, independientemente del idioma. La TABLA X al final del capítulo lo demuestra para el idioma alemán.

Como bien lo plantea George Miller [26] en la introducción de *The Psycho-Biology of Language* de la reedición de 1968, frente a esta regularidad estadística tan marcada, se puede pensar o bien que existe una propiedad universal del pensamiento humano, o que ella solo representa una consecuencia necesaria de las leyes de la probabilidad. Zipf se inclinó por trabajar sobre la primera hipótesis desarrollando la teoria del "principio del menor esfuerzo", mediante el cual pretendió explicar el aparente equilibrio entre uniformidad y diversidad en el uso del lenguaje. Sostenía que los humanos, cuando hablamos, tendemos a usar con mucha frecuencia unas pocas palabras (uniformidad), que además son de corta longitud; y por el contario, con muy poca frecuencia una gran cantidad (diversidad), de una mayor longitud. Según él, esto pone en evidencia que existe un equilibrio fundamental entre la forma y la función de los hábitos del habla, o patrones de habla en cualquier lenguaje. En 1949 aparece *"Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology"* su obra más conocida, prácticamente dedicada a justificar lo antes expuesto.

*Z*ipf era un lingüista, no un matemático, y poseía escasos conocimientos de estadística. La formulación matemática de su ley y su justificación estadística ha recibido en el último medio siglo numerosas críticas y revisiones. Sus argumentaciones no cuantitativas, basadas en principios psicológicos más que en la lógica matemática, igualmente resultaron de interés para estudiosos posteriores. Así en 1954, B. Mandelbrot, basándose en la teoría de la información y la teoría de fractales, hace un nuevo análisis de la distribución rango/frecuencia del uso del lenguaje. Diversos trabajos citan a la "ley de Zipf-Mandelbrot" como si se tratara de una sola formalización [35], [36].

CHINESE OF PEIPING

| Number of Occurrences | Number of Words | Number of Words with their Syllables | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 2045 | (315* | 1571* | 144* | 14* | 1* | 1*) |
| 2 | 405 | 110* | 358* | 43* | 3*) | | |
| 3 | 216 | 50* | 147* | 9* | 8*) | | |
| 4 | 100 | 24* | 73* | 3*) | | | |
| 5 | 69 | 29* | 38* | 2*) | | | |
| 6 | 46 | 14* | 31* | 1*) | | | |
| 7 | 41 | 10* | 25*) | | | | |
| 8 | 25 | 10* | 14* | 1*) | | | |
| 9 | 20 | 13* | 15* | 2* | 2*) | | |
| 10 | 22 | 13* | 9*) | | | | |
| 11 | 22 | 14* | 8*) | | | | |
| 12 | 22 | 15* | 7*) | | | | |
| 13 | 20 | 9* | 4*) | | | | |
| 14 | 14 | 7* | 7*) | | | | |
| 15 | 13 | 8* | 5*) | | | | |
| 16 | 10 | 4* | 5* | 1*) | | | |
| 17 | 10 | 6* | 4*) | | | | |
| 18 | 6 | 2* | 4*) | | | | |
| 19 | 5 | 4* | 1*) | | | | |
| 20 | 5 | 5*) | | | | | |
| 21 | 4 | 3* | 1*) | | | | |
| 22 | 2 | 2*) | | | | | |
| 23 | 5 | 2* | 3*) | | | | |
| 24 | 3 | 2* | 1*) | | | | |
| 28 | 4 | 3* | 1*) | | | | |
| 29 | 4 | 1* | 3*) | | | | |
| 30 | 5 | 4* | 1* | | | | |
| 31 | 6 | 4* | 2*) | | | | |
| 33 | 2 | 1* | 1*) | | | | |
| 34 | 1 | 1*) | | | | | |
| 35 | 1 | 1*) | | | | | |
| 36 | 1 | 1*) | | | | | |
| 37 | 1 | 1*) | | | | | |
| 38 | 1 | 1*) | | | | | |
| 41 | 4 | 4*) | | | | | |
| 43 | 2 | 2*) | | | | | |
| 44 | 3 | 1* | 1*) | | | | |
| 45 | 3 | 1* | 2*) | | | | |
| 47 | 2 | 2*) | | | | | |
| 50 | 1 | 1*) | | | | | |
| 52 | 1 | 1*) | | | | | |
| 55 | 2 | 2*) | | | | | |
| 57 | 1 | 1*) | | | | | |
| 58 | 1 | 1*) | | | | | |
| 59 | 1 | 1*) | | | | | |
| 66 | 2 | 1* | 1*) | | | | |
| 68 | 1 | 1*) | | | | | |
| 74 | 1 | 1*) | | | | | |
| 75 | 1 | 1*) | | | | | |
| 76 | 1 | 1*) | | | | | |
| 78 | 1 | 1*) | | | | | |
| 81 | 1 | 1*) | | | | | |
| 83 | 1 | 1*) | | | | | |
| 102 | 1 | 1*) | | | | | |
| 102–975 | 11 | ( 12*) | | | | | |
| 13,248 | 3,313 | | | | | | |

TABLA II : Zipf – Idioma Chino

LATIN OF PLAUTUS

| Number of Occurrences | Number of Words | Average Number of Syllables | Number of Occurrences | Number of Words | Average Number of Syllables |
|---|---|---|---|---|---|
| 1 | 5429 | (3.23) | 31 | 8 | |
| 2 | 1198 | (2.92) | 32 | 3 | |
| 3 | 493 | (2.77) | 33 | 4 | |
| 4 | 299 | (2.65) | 34 | 6 | (2.05) |
| 5 | 161 | (2.60) | 35 | 3 | |
| 6 | 126 | (2.53) | 36 | 5 | |
| 7 | 87 | (2.39) | 37 | 7 | |
| 8 | 69 | (2.44) | 38 | 2 | |
| 9 | 54 | (2.35) | 39 | 4 | |
| 10 | 43 | (2.33) | 40 | 3 | |
| 11 | 44 | (2.29) | 41 | 3 | |
| 12 | 36 | (2.30) | 43 | 4 | |
| 13 | 33 | (2.30) | 44 | 1 | |
| 14 | 31 | (2.09) | 45 | 1 | |
| 15 | 13 | (2.07) | 46 | 1 | |
| 16 | 25 | (2.40) | 47 | 3 | (1.70) |
| 17 | 21 | (2.09) | 48 | 1 | |
| 18 | 21 | (2.04) | 49 | 1 | |
| 19 | 11 | (2.18) | 50 | 2 | |
| 20 | 15 | | 51 | 2 | |
| 21 | 10 | | 53 | 4 | |
| 22 | 8 | (2.08) | 54 | 1 | |
| 23 | 8 | | 55 | 1 | |
| 24 | 9 | | 56 | 2 | |
| 25 | 11 | | 58 | 1 | |
| 26 | 7 | | 61 | 1 | |
| 27 | 9 | (2.00) | 62–514 | 71 | (1.40) |
| 28 | 12 | | 33,094 | 8,437 | |
| 29 | 4 | | | | |
| 30 | 4 | | | | |

TABLA III: Zipf – Idioma Latín

AMERICAN NEWSPAPER ENGLISH
(According to R. C. Eldridge)

| Number of Occurrences | Number of Words | Average Number of Phonemes |
|---|---|---|
| 1 | 2976 | (6.656) |
| 2 | 1079 | (6.151) |
| 3 | 516 | (6.015) |
| 4 | 294 | (6.081) |
| 5 | 212 | (5.589) |
| 6 | 151 | (5.768) |
| 7 | 105 | (5.333) |
| 8 | 84 | (5.654) |
| 9 | 86 | (5.174) |
| 10 | 45 | (5.377) |
| 11 | 40 | (4.825) |
| 12 | 37 | (5.459) |
| 13 | 25 | (5.560) |
| 14 | 28 | (5.00) |
| 15 | 26 | (4.807) |
| 16 | 17 | (5.058) |
| 17 | 18 | (4.166) |
| 18 | 10 | (6.100) |
| 19 | 15 | (4.733) |
| 20 | 16 | (4.687) |
| 21 | 13 | |
| 22 | 11 | |
| 23 | 6 | |
| 24 | 8 | |
| 25 | 6 | (3.455) |
| 26 | 10 | |
| 27 | 9 | |
| 28 | 6 | |
| 29 | 5 | |
| 30 | 4 | |
| 31 | 6 | |
| 32 | 4 | |
| 33 | 6 | |
| 34 | 2 | |
| 35 | 5 | |
| 36 | 3 | |
| 37 | 2 | |
| 39 | 2 | |
| 40 | 4 | |
| 41 | 1 | (3.903) |
| 42 | 7 | |
| 43 | 1 | |
| 44 | 4 | |
| 45 | 1 | |
| 46 | 2 | |
| 47 | 5 | |
| 48 | 1 | |
| 49 | 3 | |
| 50 | 3 | |
| 51 | 1 | |
| 52 | 3 | |
| 54 | 1 | |
| 55 | 1 | (3.333) |
| 56 | 1 | |
| 58 | 2 | |
| 60 | 1 | |
| 61–4290 | 71 | (2.666) |

TABLA IV: Zipf – Idioma Inglés

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 4 | palabras | que | ocurren | 30 | veces | $30^2 = 900$ | 2976/900= | 3,3 |
| 5 | ” | ” | ” | 29 | “ | $29^2 = 841$ | 2976/841= | 3,5 |
| 6 | ” | ” | ” | 28 | “ | ..... | | |
| 9 | ” | ” | ” | 27 | “ | ..... | | |
| 10 | “ | “ | “ | 26 | “ | ..... | | |
| 6 | ” | ” | ” | 25 | “ | ..... | | |
| 8 | ” | ” | ” | 24 | “ | $24^2 = 576$ | 2976/576= | 5,1 |
| 6 | ” | ” | ” | 23 | “ | ..... | | |
| 11 | “ | “ | “ | 22 | “ | ..... | | |
| 13 | “ | “ | “ | 21 | “ | ..... | | |
| 16 | “ | “ | “ | 20 | “ | ..... | | |
| 15 | “ | “ | “ | 19 | “ | ..... | | |
| 10 | “ | “ | “ | 18 | “ | ..... | | |
| 18 | “ | “ | “ | 17 | “ | $17^2 = 289$ | 2976/289= | 10,2 |
| 17 | “ | “ | “ | 16 | “ | ..... | | |
| 26 | “ | “ | “ | 15 | “ | ..... | | |
| 28 | “ | “ | “ | 14 | “ | ..... | | |
| 25 | “ | “ | “ | 13 | “ | ..... | | |
| 37 | “ | “ | “ | 12 | “ | $12^2 = 144$ | 2976/144= | 20,6 |
| 40 | “ | “ | “ | 11 | “ | ..... | | |
| 45 | “ | “ | “ | 10 | “ | ..... | | |
| 86 | “ | “ | “ | 9 | “ | ..... | | |
| 84 | “ | “ | “ | 8 | “ | ..... | | |
| 105 | ” | ” | ” | 7 | “ | $6^2 = 36$ | 2976/36= | 82,6 |
| 151 | ” | ” | ” | 6 | “ | ..... | | |
| 212 | ” | ” | ” | 5 | “ | ..... | | |
| 294 | ” | ” | ” | 4 | “ | $3^2 = 9$ | 2976/9= | 330,6 |
| 516 | ” | ” | ” | 3 | “ | $2^2 = 4$ | 2976/4= | 744 |
| 1079 | ” | ” | ” | 2 | “ | | | |
| X=2976 palabras que ocurren 1 vez | | | | | | | | |

TABLA V: Correlación entre la frecuencia de una palabra y el número de palabras que ocurren con dicha frecuencia (Fuente de datos Tabla IV: Zipf – Idioma Inglés).

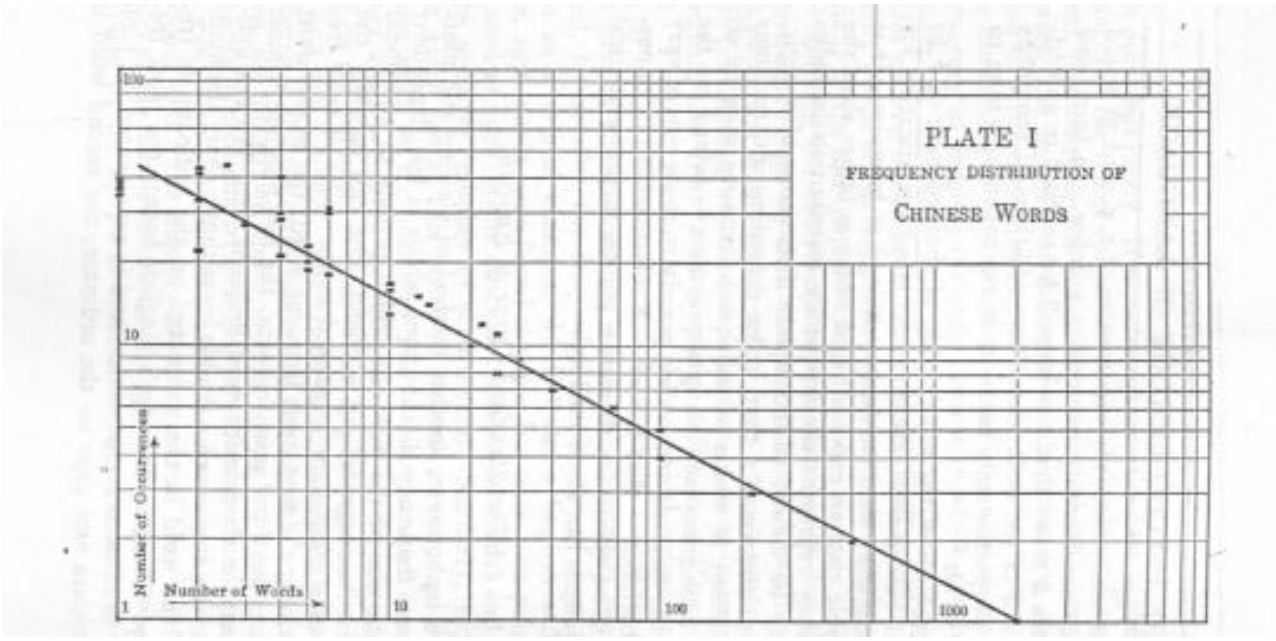



GRAFICO II: Correlación entre la frecuencia de las palabras en iidioma chino y el número de palabras que ocurren con dicha frecuencia graficado en escala logarítmica.

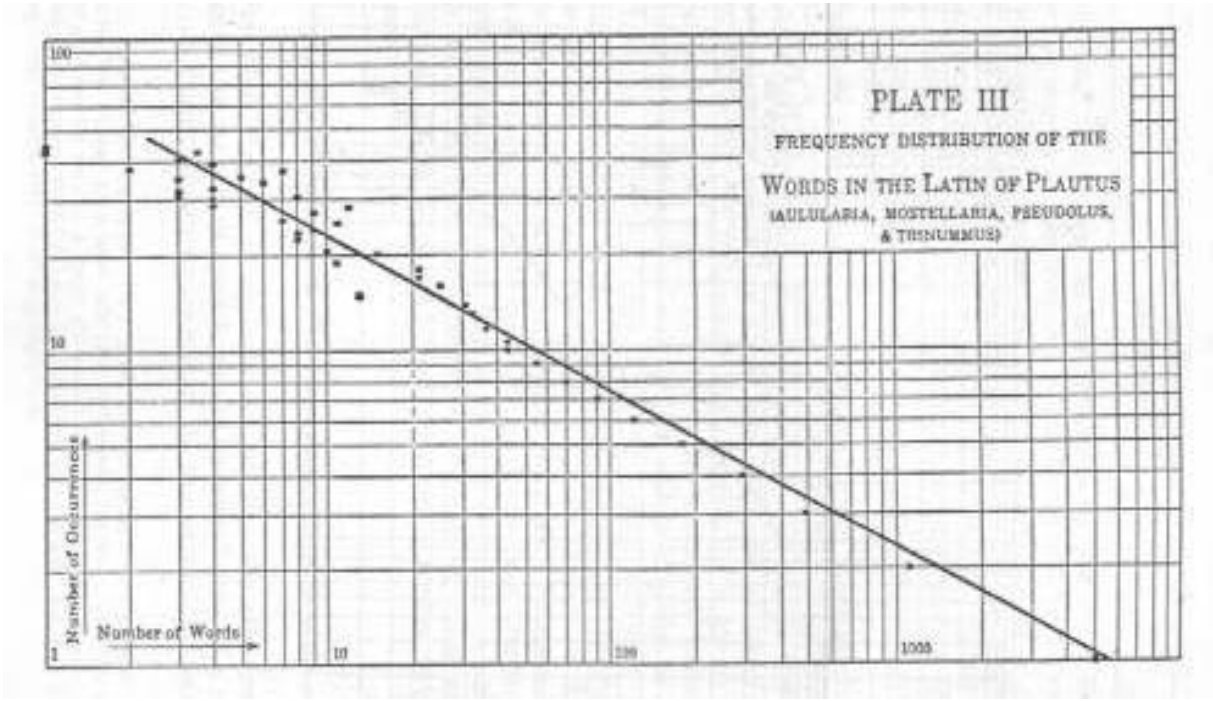



GRAFICO III: Correlación entre la frecuencia de las palabras en idioma latín y el número de palabras que ocurren con dicha frecuencia graficado en escala logarítmica.

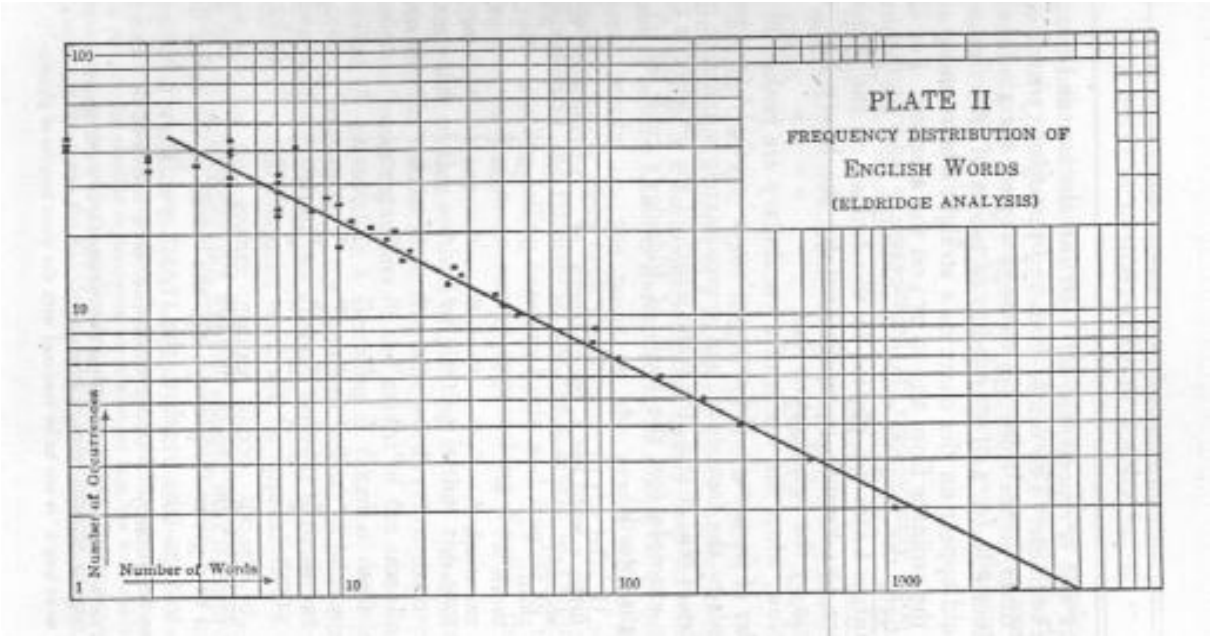



GRAFICO IV: Correlación entre la frecuencia de las palabras en idioma inglés y el número de palabras que ocurren con dicha frecuencia graficado en escala logarítmica.

TABLA II: Chino

| *n* | $n^2$ | *f* | $(f * n^2)$ |
|---|---|---|---|
| 1 | 1 | 2046 | 2046 |
| 2 | 4 | 494 | 1976 |
| 3 | 9 | 216 | 1944 |
| 4 | 16 | 100 | 1600 |
| 5 | 25 | 99 | 2475 |
| 6 | 36 | 66 | 2376 |
| 7 | 49 | 41 | 2009 |
| 8 | 64 | 25 | 1600 |
| 9 | 81 | 30 | 2430 |
| 10 | 100 | 20 | 2000 |
| 11 | 121 | 25 | 3025 |
| 12 | 144 | 22 | 3168 |
| 13 | 169 | 10 | 1690 |
| 14 | 196 | 14 | 2744 |
| 15 | 225 | 13 | 2925 |
| 16 | 256 | 10 | 2560 |
| 17 | 289 | 10 | 2890 |
| 18 | 324 | 6 | 1944 |
| 19 | 361 | 5 | 1805 |
| 20 | 400 | 5 | 2000 |
| 21 | 441 | 4 | 1764 |
| 22 | 484 | 2 | 968 |
| 23 | 529 | 5 | 2645 |
| 26 | 676 | 3 | 2028 |
| 28 | 784 | 4 | 3136 |
| 29 | 841 | 4 | 3364 |
| 30 | 900 | 6 | 5400 |
| 32 | 1024 | 6 | 6144 |
| 33 | 1089 | 2 | 2178 |
| 34 | 1156 | 1 | 1156 |
| 35 | 1225 | 1 | 1225 |
| 36 | 1296 | 1 | 1296 |
| 37 | 1369 | 1 | 1369 |
| 38 | 1444 | 1 | 1444 |
| 41 | 1681 | 4 | 6724 |
| 43 | 1849 | 2 | 3698 |
| 44 | 1936 | 2 | 3872 |
| 45 | 2025 | 3 | 6075 |
| 46 | 2116 | 1 | 2116 |
| 47 | 2209 | 2 | 4418 |
| 50 | 2500 | 1 | 2500 |
| 52 | 2704 | 1 | 2704 |
| 55 | 3025 | 2 | 6050 |
| 57 | 3249 | 1 | 3249 |
| 58 | 3364 | 1 | 3364 |
| 60 | 3600 | 1 | 3600 |
| 66 | 4356 | 2 | 8712 |
| 68 | 4624 | 1 | 4624 |
| 72 | 5184 | 1 | 5184 |
| 73 | 5329 | 1 | 5329 |
| 75 | 5625 | 1 | 5625 |
| 78 | 6084 | 1 | 6084 |
| 81 | 6561 | 1 | 6561 |
| 83 | 6889 | 1 | 6889 |
| 101 | 10201 | 2 | 20402 |

TABLA III: Latín

| *n* | $n^2$ | *f* | $(f * n^2)$ |
|---|---|---|---|
| 1 | 1 | 5429 | 5429 |
| 2 | 4 | 1198 | 4792 |
| 3 | 9 | 492 | 4428 |
| 4 | 16 | 299 | 4784 |
| 5 | 25 | 161 | 4025 |
| 6 | 36 | 126 | 4536 |
| 7 | 49 | 87 | 4263 |
| 8 | 64 | 69 | 4416 |
| 9 | 81 | 54 | 4374 |
| 10 | 100 | 43 | 4300 |
| 11 | 121 | 44 | 5324 |
| 12 | 144 | 36 | 5184 |
| 13 | 169 | 33 | 5577 |
| 14 | 196 | 31 | 6076 |
| 15 | 225 | 13 | 2925 |
| 16 | 256 | 25 | 6400 |
| 17 | 289 | 21 | 6069 |
| 18 | 324 | 21 | 6804 |
| 19 | 361 | 11 | 3971 |
| 20 | 400 | 15 | 6000 |
| 21 | 441 | 10 | 4410 |
| 22 | 484 | 8 | 3872 |
| 23 | 529 | 8 | 4232 |
| 24 | 576 | 9 | 5184 |
| 25 | 625 | 11 | 6875 |
| 26 | 676 | 7 | 4732 |
| 27 | 729 | 9 | 6561 |
| 28 | 784 | 12 | 9408 |
| 29 | 841 | 4 | 3364 |
| 30 | 900 | 4 | 3600 |
| 31 | 961 | 8 | 7688 |
| 32 | 1024 | 3 | 3072 |
| 33 | 1089 | 4 | 4356 |
| 34 | 1156 | 6 | 6936 |
| 35 | 1225 | 3 | 3675 |
| 36 | 1296 | 5 | 6480 |
| 37 | 1369 | 7 | 9583 |
| 38 | 1444 | 2 | 2888 |
| 39 | 1521 | 4 | 6084 |
| 40 | 1600 | 3 | 4800 |
| 41 | 1681 | 3 | 5043 |
| 43 | 1849 | 4 | 7396 |
| 44 | 1936 | 1 | 1936 |
| 45 | 2025 | 1 | 2025 |
| 46 | 2116 | 1 | 2116 |
| 47 | 2209 | 3 | 6627 |
| 48 | 2304 | 1 | 2304 |
| 49 | 2401 | 1 | 2401 |
| 50 | 2500 | 2 | 5000 |
| 51 | 2601 | 2 | 5202 |
| 53 | 2809 | 4 | 11236 |
| 54 | 2916 | 1 | 2916 |
| 55 | 3025 | 1 | 3025 |
| 56 | 3136 | 2 | 6272 |
| 58 | 3364 | 1 | 3364 |
| 61 | 3721 | 3 | 11163 |

TABLA IV: Inglés

| *n* | $n^2$ | *f* | $(f * n^2)$ |
|---|---|---|---|
| 1 | 1 | 2976 | 2976 |
| 2 | 4 | 1079 | 4316 |
| 3 | 9 | 516 | 4644 |
| 4 | 16 | 294 | 4704 |
| 5 | 25 | 212 | 5300 |
| 6 | 36 | 151 | 5436 |
| 7 | 49 | 105 | 5145 |
| 8 | 64 | 84 | 5376 |
| 9 | 81 | 86 | 6966 |
| 10 | 100 | 45 | 4500 |
| 11 | 121 | 40 | 4840 |
| 12 | 144 | 37 | 5328 |
| 13 | 169 | 25 | 4225 |
| 14 | 196 | 28 | 5488 |
| 15 | 225 | 26 | 5850 |
| 16 | 256 | 17 | 4352 |
| 17 | 289 | 18 | 5202 |
| 18 | 324 | 10 | 3240 |
| 19 | 361 | 15 | 5415 |
| 20 | 400 | 16 | 6400 |
| 21 | 441 | 13 | 5733 |
| 22 | 484 | 11 | 5324 |
| 23 | 529 | 6 | 3174 |
| 24 | 576 | 8 | 4608 |
| 25 | 625 | 6 | 3750 |
| 26 | 676 | 10 | 6760 |
| 27 | 729 | 9 | 6561 |
| 28 | 784 | 6 | 4704 |
| 29 | 841 | 5 | 4205 |
| 30 | 900 | 4 | 3600 |
| 31 | 961 | 6 | 5766 |
| 32 | 1024 | 4 | 4096 |
| 33 | 1089 | 6 | 6534 |
| 34 | 1156 | 2 | 2312 |
| 35 | 1225 | 5 | 6125 |
| 36 | 1296 | 3 | 3888 |
| 37 | 1369 | 2 | 2738 |
| 39 | 1521 | 2 | 3042 |
| 40 | 1600 | 4 | 6400 |
| 41 | 1681 | 1 | 1681 |
| 42 | 1764 | 7 | 12348 |
| 43 | 1849 | 1 | 1849 |
| 44 | 1936 | 4 | 7744 |
| 45 | 2025 | 1 | 2025 |
| 46 | 2116 | 2 | 4232 |
| 47 | 2209 | 5 | 11045 |
| 48 | 2304 | 1 | 2304 |
| 49 | 2401 | 3 | 7203 |
| 50 | 2500 | 3 | 7500 |
| 51 | 2601 | 1 | 2601 |
| 52 | 2704 | 3 | 8112 |
| 54 | 2916 | 1 | 2916 |
| 55 | 3025 | 1 | 3025 |
| 56 | 3136 | 1 | 3136 |
| 58 | 3364 | 2 | 6728 |
| 60 | 3600 | 1 | 3600 |

TABLA VI: cálculo de la constante resultante de $n * f^2$, para los valores de las muestras consignadas en las TABLAS II, III y IV

| Nro.sílabas En la palabra | Nro.ocurrencias (incluye repeticiones) | Porcentaje del total |
|---|---|---|
| 1 | 5.426.326 | 49,76% |
| 2 | 3.156.448 | 28,94% |
| 3 | 1.410.494 | 12,93% |
| 4 | 646.971 | 5,93% |
| 5 | 187.738 | 1,72% |
| 6 | 54.436 | 0,50% |
| 7 | 16.993 | |
| 8 | 5.038 | |
| 9 | 1.225 | |
| 10 | 461 | |
| 11 | 59 | 0,22% |
| 12 | 35 | |
| 13 | 8 | |
| 14 | 2 | |
| 15 | 1 | |
| | 10.906.235 | 100.00% |

TABLA X: Extraída de *"The psycho-biology of language"* / G.K. Zipf. [32] p.23

# Capítulo 4

## Antecedentes del tratamiento automático de textos: los primeros tiempos de la computación

En su trabajo *"The seven ages of information retrieval",* M. Lesk [37] sostiene que la primera etapa en la evolución de la RI se dio en el periodo 1945-55. En esos años V. Bush[1] expuso las ideas que posteriormente inspirarían al hipertexto; W. Wiener[2] realizó los trabajos sobre encriptación que darían lugar a las investigaciones sobre traducción automática; H.P. Luhn[3] aplicó el concepto de concordancia en el desarrollo de los índices KWIC; y C. Mooers[4] desarrolló el sistema Zatocoding de tarjetas perforadas. Luego, en los años siguientes, se produciría una etapa de fuerte experimentación que duraría hasta finales de los 60, durante la cual, el proyecto Cranfield marcó un hito fundamental. Con él se produjo el nacimiento de la evaluación en la recuperación de información al desarrollar C. Cleverdon[5] las clásicas medidas de exhaustividad y precisión. Este mismo proyecto constituyó el antecedente del uso de las colecciones de prueba como medio de sistematización de los experimentos que luego serán implementadas con reconocida efectividad en el marco de las Conferencias TREC[6].

Durante esta segunda etapa tuvieron lugar algunos trabajos que influyeron especialmente en el desarrollo de las técnicas de indización automática. Se hace referencia aquí a algunas contribuciones, considerando que el conteo de frecuencias medias de H.P. Luhn, la visión de la RI desde una perspectiva probabilística de Maron y Kuhns, y la co-ocurrencia de palabras propuestas por Maron [et al.], Stiles y Doyle; son las más significativas.

### 4.1. H.P. Luhn: la frecuencia de las palabras y su valor discriminante

[1] BUSH, V., ref.3, p.7
[2] WEAVER, W. Translation. En *Machine Translation of Language.* Locke, W.N.; Booth, A.D. New York: Wiley, 1955.
[3] LUHN, H.P. The International Conference on Scientific Information (ICSI). Washington, 1958.
[4] MOOERS, C.N. Zatocoding and Development in Information Retrieval. En *ASLIB Proc*. February 1956, p. 3
[5] CLEVERDON, C. *Proposals for an investigation into the efficiency of various retrieval systems*. Research proposal. Cranfield University, 1956.
[6] TREC, ref.8, p.10

Hans Peter Luhn (1896-1964) fue un precursor indiscutible de los trabajos relacionados con la automatización en los procesos de recuperación de información. Su interés en esta área, comenzó en los últimos veinte años de su vida, cuando siendo empleado de IBM y preocupado por la "explosión de la información", aplicó su natural inventiva al empleo de máquinas para procesar y recuperar la literatura. Dentro de los principales aportes que se le adjudican en la Ciencia de la Información se pueden mencionar: los índices permutados, la generación automática de resúmenes y la diseminación selectiva de la información (SDI). También fue el primero en concebir un Sistema de Recuperación de Información basado en el modelo del espacio vectorial[7], aporte que no siempre es considerado en la bibliografía especializada.

En el presente apartado se hace referencia a la contribución que se considera más pertinente a los fines de este trabajo que es su desarrollo sobre la generación automática de resúmenes, en el cual postula la idea del **valor de discriminación** de las palabras con frecuencias intermedias. La importancia de este concepto es fundamental en las técnicas de indización automática de enfoque estadístico que se desarrollaron posteriormente.

En 1958, H.P. Luhn, publica el tabajo "*The automatic creation of literature abstracts"* [38]*.* En dicho trabajo, describe una investigación exploratoria en métodos automáticos de obtención de resúmenes. El sistema descrito permitía seleccionar de todas las oraciones del texto, aquellas que eran más significativas o representativas de la información que éste contenía. Este programa podía ser aplicado únicamente a literatura específica en ciencia y tecnología. Para determinar que oraciones del artículo podían servir como "auto-resumen", se requería una medida por la cual la información contenida en la totalidad de las oraciones del texto fuera comparada. Así, del análisis de las palabras se derivó el "factor de significación" a partir de:

1) la frecuencia de las palabras del artículo podía ser usada para representar diferentes grados de significado.
2) la posición relativa en la oración de ciertas palabras dadas, podía ser usada como forma de medir la significación de las oraciones.

La justificación de medir la significación de las palabras basándose en su frecuencia de aparición se basa en que, según Luhn, normalmente la persona que escribe repite ciertas palabras para argumentar sus ideas. Sostiene que existe una muy baja probabilidad de que una palabra dada refleje más de una noción, así como

[7] LUHN, H.P. A new method of recording and searching information*. American Documentation*. 1953, vol.4, p.14-16.

también, existe una muy baja probabilidad de que un autor utilice diferentes palabras para reflejar una misma noción (se debe pensar aquí que se está hablando de la terminología científico-tecnológica). Aún, cuando el autor por razones estilísticas emplea la sinonimia, corre en busca de alternativas legítimas y cae en repeticiones al reforzar su idea principal. Partiendo de estos argumentos, Luhn toma la curva de Zipf, que como ya se explicó en el apartado anterior muestra la distribución de palabras en un texto basándose en la frecuencia; y presenta la zona que él consideraría como la de las palabras significativas (GRAFICO I):

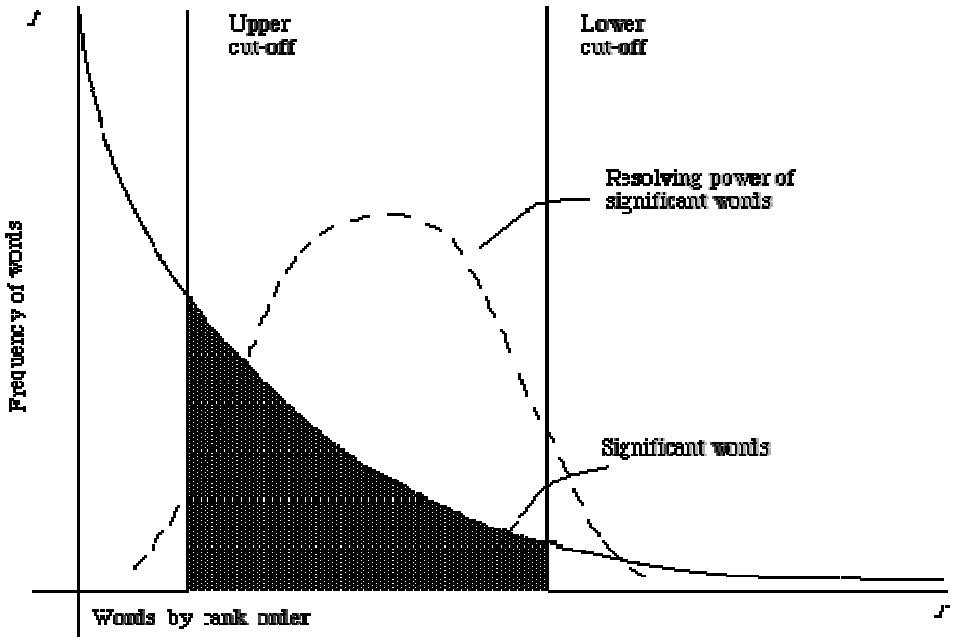



*Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz[44] page 120)*

GRAFICO I: Curva de Luhn

En la región de las altas frecuencias se dan muchas palabras comunes que solo constituyen “ruido” en el sistema. Este ruido se reduce aplicando una técnica de eliminación, que consiste en comparar las palabras del texto con una lista previamente compilada de palabras no-significativas. La forma de determinar las líneas de corte entre las palabras de alta frecuencia y las de baja (líneas C y D de la gráfica), debe basarse en la experiencia con varios ejemplos de trabajos publicados. Las palabras comprendidas entre ambas líneas de corte describen una curva que marca el grado de discriminación o poder de resolución de las mismas (línea E).

Luhn no hace consideraciones acerca del significado de las palabras o de los argumentos expresados por las combinaciones de palabras. Argumenta que,

cualquiera sea el tópico, se tiende a asociar ciertas palabras que son específicas en un aspecto del tema que esta siendo tratado. Además, si las palabras que tienen una frecuencia alta (recordemos que las de frecuencia “muy“ alta ya fueron eliminadas) se encuentran en una gran proximidad física unas de otras, entonces existe una probabilidad muy alta de que la información allí involucrada sea la más representativa del artículo. Tanto en el lenguaje hablado como en el escrito, las ideas que están más fuertemente asociadas desde el punto de vista intelectual, se manifiestan a través de ciertas asociaciones de palabras que se presentan físicamente juntas. Al respecto, en 1957, había publicado el trabajo *“A statistical approach to mechanized encoding and searching of literary information”* [39]. En dicho trabajo presentó la noción de que la comunicación de ideas por medio de palabras tiene una base probabilística. Fundamentó su afirmación sosteniendo que cuando un emisor quiere comunicarle algo a alguien, realiza una selección de palabras por medio de las cuales pretende lograr en la mente del receptor un estado similar al de su propia mente. Con la finalidad de una mejor comunicación, se divide la idea en una serie de ideas más pequeñas hasta que se siente que se ha llegado a un nivel de nociones compartidas cuya prolongación dependerá del grado de similitud de experiencias comunes. Cuanta menos experiencia tengan en común, mayor cantidad de palabras se deberá usar. Cuando se refiere a la comunicación escrita este proceso adquiere características estáticas. El emisor supone como será su potencial receptor y selecciona un cierto nivel de experiencias, necesario para que exista la comunicación, tomando como marco de referencia lo que otros autores han escrito sobre el tema en cuestión y que se supone ha servido para elevar el nivel de experiencia común. Bajo estas consideraciones es que Luhn propone calcular el “factor de significación” de las oraciones de la siguiente manera:

1) Tomar solo de las oraciones las partes donde se encuentran concentradas las palabras significativas y formar con ellas agrupamientos (recordar que las palabras significativas se determinan por el conteo de frecuencia). Luhn sostiene que por pruebas realizadas, cada agrupamiento o “cluster” no debería contener más de 4 ó 5 palabras no significativas entre medio de palabras significativas. Es decir que para una oración se puede determinar más de un agrupamiento.
2) Contar dentro de cada agrupamiento el número de palabras significativas y dividir el cuadrado de ese número por el número total de palabras dentro del agrupamiento.
3) Tomando el número obtenido, se ordenan las oraciones de mayor a menor. Si el texto es muy extenso, a los fines de confeccionar un mejor “auto-resumen” es

conveniente tomar las oraciones con factor de significación más alto por sección, en lugar de hacerlo globalmente.

## 4.2. Maron y Kuhns: la probabilidad de relevancia

En un capítulo anterior se ha mencionado el modelo de Recuperación de Información con base probabilística y se ha citado como una de las primeras investigaciones al trabajo de Maron y Kuhns [18]. En dicho trabajo, los autores proponen una técnica de indización y recuperación, en la cual, frente a una interrogación dada, el sistema realiza una inferencia estadística para calcular un número que cuantifique la **relevancia** de cada documento para la interrogación. Luego, mediante comparaciones sucesivas, el sistema es capaz de realizar un ordenamiento en la salida. Este concepto de relevancia se basa en un cálculo de la probabilidad de que el documento satisfaga el requerimiento de información.

El problema de cuantificar un concepto primario como es el de relevancia no es trivial. Maron y Kuhns tomaron como punto de partida para su desarrollo la idea de que en cierto sentido, el problema de explicar la noción de relevancia, concepto básico en la teoría de la Recuperación de Información, es similar a explicar la noción de **cantidad de información**, concepto básico en la Teoría de la Comunicación. Es por ello que, primeramente, se hace referencia en el apartado siguiente a los conceptos básicos de la Teoría de la Información de Shannon y Weaver[8], para luego introducir lo expuesto por Maron y Kuhn [18] y Maron [40].

### 4.2.1 Shannon y Weaver: la medida de cantidad de información

Se expondrá en el presente apartado como la Teoría de la Información trata de encontrar la medida de cantidad de información en término de probabilidades. Se parte de considerar ejemplos simples, donde $R_0$ es la cantidad de eventos diferentes posibles que son igualmente probables a priori. Por ejemplo, en el caso del lanzamiento de la moneda, los eventos cara o cruz, nos determina que $R_0=2$. En el caso del dado, se tiene que $R_0=6$. Tanto el lanzamiento de la moneda como del dado, puede interpretarse como la recepción de un mensaje, donde una sola posibilidad de $R_0$ se da. Aparentemente, cuanto más grande es $R_0$ mayor incertidumbre hay antes de la recepción del mensaje y más grande será la cantidad de información después de que el mensaje fue recibido. Se puede interpretar esto de la siguiente manera: en la situación inicial no se tiene información, por ejemplo $I_0 = 0$ con $R_0$ eventos igualmente

[8] SHANNON, C. y WEAVER, W. *The mathematical theory of communication*. Illinois, 1949.

probables. En el estado final, se tiene información $I_1 \neq 1$ con $R_1 = 1$, un solo evento (por ejemplo una tirada de dado). Se deduce entonces que la cantidad de información está aparentemente conectada de alguna manera con la cantidad de eventos distintos posibles. Para tener una idea de cómo es esta conexión, se debe entender que *I* es aditiva para eventos independientes. Para dos eventos, $R_{01}$ y $R_{02}$, la cantidad de información es:

$$I(R_{01}R_{02}) = I(R_{01}) + I(R_{02})$$

Si se imagina una fuente de información que emite una secuencia de símbolos pertenecientes a un alfabeto finito y fijo, $R_0 = \{ R_{01}, R_{02}, .. R_{0q}\}$ donde cada símbolo emitido se elige de acuerdo con una ley fija de probabilidad. Este tipo de fuente de información sencilla se conoce como fuente de información de memoria nula y puede describirse completamente mediante el alfabeto $R_0$ y las probabilidades con que los símbolos aparecen:

$$\mathrm{P}(\mathrm{R}_{01}), \mathrm{P}(\mathrm{R}_{02}),..., \mathrm{P}(\mathrm{R}_{\mathrm{q}})$$

Se define a la **medida de información** como:

$$I(R_{0i}) = \log \frac{1}{P(R_{0i})}$$

Dependiendo de la base del logaritmo que se elija se obtiene la medida en diferentes unidades. Así, si se elige el logaritmo en base 2, la cantidad de información se obtiene en **bits**. Si se emplea logaritmos naturales la unidad es el **nat**. Si se emplea un logaritmo en base 10 la unidad es el **Hartley**. La unidad elegida en la presente exposición es el bit por considerarla la de uso más común.

Claramente se observa que a mayor probabilidad de aparición del símbolo, menor será la cantidad de información recibida. La cantidad media de bits de información por símbolo de la fuente es

$$\frac{\sum I(R_i) * F(R_i)}{\sum F(R_i)} = \frac{\sum I(R_i) * F(R_i)}{n} =$$

$$= \sum I(R_i) * \frac{F(R_i)}{n} = \sum_{R_0} I(R_i) P(R_i)$$

Esta cantidad media de información por símbolo de la fuente recibe el nombre de *entropía H(S)* de la fuente de memoria nula [41, p.28 ].

$$H(S) = \sum_{R_0} P(R_i) * \log \frac{1}{P(R_i)}$$

### 4.2.3 La medida de cantidad de relevancia

En un trabajo de 1961, Maron [40] propone un experimento de indización automática en el cual refleja los conceptos antes expuestos. El experimento consistía en tomar un conjunto de documentos y clasificarlos manualmente con un conjunto de categorías temáticas. Luego se seleccionaba un conjunto de palabras significativas nativas de los documentos basándose en la frecuencia, y una vez que los documentos habían sido ordenados en sus respectivas categorías se determinaba una correlación estadística entre las palabras y dichas categorías. Es decir que se obtenía una lista que mostraba el número de veces que cada una de las diferentes palabras seleccionadas provenían de documentos indizados con la categoría 1, la categoría 2, etc. Luego se tomaba otro corpus textual y se lo indizaba automáticamente teniendo como base la información estadística anterior. La idea era tomar cada nuevo documento y que el sistema contrastara automáticamente que palabras del listado anterior estaban presentes en el documento y basándose en la frecuencia, realizara una inferencia de la categoría temática a la que pertenece.

La base teórica que se utiliza para realizar la inferencia es que si se toma cualquier palabra del documento, existe un nivel de incertidumbre sobre la categoría a la cual el documento en cuestión pertenece. Esta incertidumbre es representada por la distribución probabilística de las categorías y según Maron puede ser medida por la expresión de la entropía de Shannon. Sea la probabilidad de que un documento cualquiera sea indizado bajo una categoría $P(C_j)$, se tiene que la incertidumbre es:

$$H = -\sum_{j=1}^{n} P(C_j) * \log_2 P(C_j)$$

Si una palabra particular *W* ocurre en dicho documento, se tendrá entonces que la incertidumbre es:

$$H' = -\sum_{j=1}^{n} P(C_j) \mid W_i) * \log_2 P(C_j \mid W_{ii})$$

Entonces, la cantidad de incertidumbre que puede ser eliminada está dada por la diferencia entre *H'* y *H*. Dadas dos palabras $W_1$ y $W_2$, $W_1$ será más significativa si su ocurrencia en el documento elimina una mayor cantidad de incertidumbre inicial que si ocurriera la palabra $W_2$. Dado un documento $d_1$ que contiene una palabra significativa $W_1$ , ¿cuál será la probabilidad de que dicho documento pertenezca a la categoría $C_1$, $C_2$, etc…?

$$P(C_j \mid W_i) = \frac{P(C_j) * P(W_{i\mid} \mid C_j)}{P(W_1)}$$

Se estima el valor $P(C_j)$ como la cantidad de palabras significativas indizadas bajo esa categoría dividido el número total de palabras significativas. El valor de $P(W_i \mid C_j)$ se estimará como la relación entre la cantidad de ocurrencias de la palabra *i ésima* que pertenece al documento que está indizando en la categoría *j ésima* y la cantidad de ocurrencias en todos los documentos que pertenecen a la categoría *j ésima*.

## 4.3. H.E. Stiles y L.B. Doyle: co-ocurrencia de palabras

En 1962 L.B. Doyle [42] publicó un trabajo donde exponía el uso de la co-ocurrencia de palabras aplicado a la indización automática. Tomando como punto de partida un informe elaborado para la compañía System Development Corporation en la cual trabajaba, Doyle manifiesta que un autor que escribe sobre temas muy especializados utiliza ciertas palabras con una frecuencia inusual, y que como consecuencia de esto también se da de manera inusual la co-ocurrencia de ciertos pares de palabras dentro del mismo texto. Cuando dos pares de palabras co-ocurren con frecuencia en un texto no debería sorprendernos que ambas palabras tengan una relación asociativa fuerte en la mente del autor. Doyle sostiene que es razonable asumir que si bien los diferentes autores dentro de una misma temática realizarán asociaciones libres de palabras, en términos estadísticos el uso del vocabulario de la especialidad, de uso recomendado para transmitir el mensaje adecuadamente, conduce a los autores a realizar asociaciones similares.

Para aplicar la co-ocurrencia, Doyle manifiesta que se puede optar por considerar que:

1) Dos palabras co-ocurren cuando por lo menos ambas palabras están presentes una vez en la unidad textual que se tome.
2) Dos palabras co-ocurren cuando por lo menos ambas tienen una frecuencia en la unidad textual mayor que la establecida por cierto valor predeterminado.
3) Dos palabras co-ocurren o no co-ocurren, hablando en términos de cantidad de co-ocurrencias la cual varía en función de la frecuencia de ambas palabras en la unidad textual.

Además introduce el concepto de **prevalencia** (que luego otros autores denominarán *frecuencia relativa* o *frecuencia global*) para diferenciar dos tipos de cantidades. Para Doyle, la **frecuencia** es la cantidad de veces que la palabra aparece en el conjunto de los textos de la colección, mientras que la **prevalencia** es el número total de textos de la colección que la contiene. Ambos valores no tienen que ser proporcionales ya que los agrupamientos (**clusters**) de palabras pueden darse sin afectar significativamente la frecuencia a nivel de la colección.

Para poder medir de alguna manera los patrones de co-ocurrencia, H.E. Stiles [43] propone la siguiente función que llamó factor de asociación:

$$\log_{10} \frac{(\left| fN - AB \right| - N/2)^2 N}{AB(N-A)(N-B)}$$

f= prevalencia de la co-ocurrencia de la palabra A y la palabra B (conjunto de documentos en los que están las dos)
N= cantidad total de documentos
A= prevalencia de la palabra A
B= prevalencia de la palabra B

Luego, H. Borko[9] desarrollaría una función similar basada en el coeficiente de correlación de Pearson[10]:

[9] BORKO, H. The construction of an empirically based mathematically derived classification system. *Proceedings of the Western Joint Computer Conference.* May, 1962.

[10] El coeficiente de Pearson es un índice que mide la relación entre dos variables cuantitativas. Su cálculo se realiza dividiendo la covarianza por el producto de las desviaciones estándar de ambas variables.

$$\frac{fN - AB}{\sqrt{AB(N - A)(N - B)}}$$

Sin embargo, Doyle sostiene que introducir la variable del tamaño de la colección ocasiona inconvenientes y brinda el siguiente ejemplo: para una colección de 100 libros de neurociencia, donde la palabra “neurona” ocurre 80 veces, la palabra “sinapsis” ocurre 75 y ambas palabras co-ocurren 58 veces, aplicando la función de Borko, se tiene un factor de asociación del -200. Esto significa que se estaría frente a una correlación negativa. Ahora bien, si se colocan esos 100 libros en una biblioteca técnica junto con otros 9900, al calcular el factor de asociación, se ve que ha variado considerablemente mostrando un índice de correlación alto entre A y B. ¿Se justifica entonces introducir el tamaño de la colección? Sí, si lo que se busca es que la asociación sea un atributo de toda la biblioteca. No, si lo que se busca es que la asociación sea un atributo de los 100 libros de neurología quienes no cambiarán su naturaleza intrínseca porque fueron adicionados a una biblioteca. Desde este punto de vista. La variable N introduce un valor no deseado en la función de cálculo de la asociación entre dos palabras. Por ello L. Doyle prefiere proponer una función que involucra solo a f, A y B:

$$\frac{f}{A + B - f}$$ [11]

Los trabajos experimentales arrojaron que la función favorece a aquellos pares de palabras en que la prevalencia es similar.

Su uso en la RI permitiría que las búsquedas no solo recuperen los documentos que contienen las palabras sino que también se puedan recuperar los documentos que están fuertemente asociados. A Doyle se le ocurrió su aplicación en la generación de mapas asociativos para el armado de interfaces de recuperación, que si bien no llegó a desarrollar, su trabajo constituye un temprano aporte a este tipo de productos que en la actualidad se consideran de avanzada. En el GRAFICO II se muestra una matriz de correlación de 90 X 90 palabras en un corpus de 618 resúmenes del Psychological Abstracts. Cada una de las 90 palabras fue correlacionada con las otras 90 mediante el índice de Pearson. En el mapa, cada par de palabras está conectado mediante enlaces que muestran el mayor o menor nivel de asociación. El número sobre los enlaces es el valor del coeficiente de correlación.

[11] Esta ecuación plantea el interrogante si no debería ser A+B-2f?

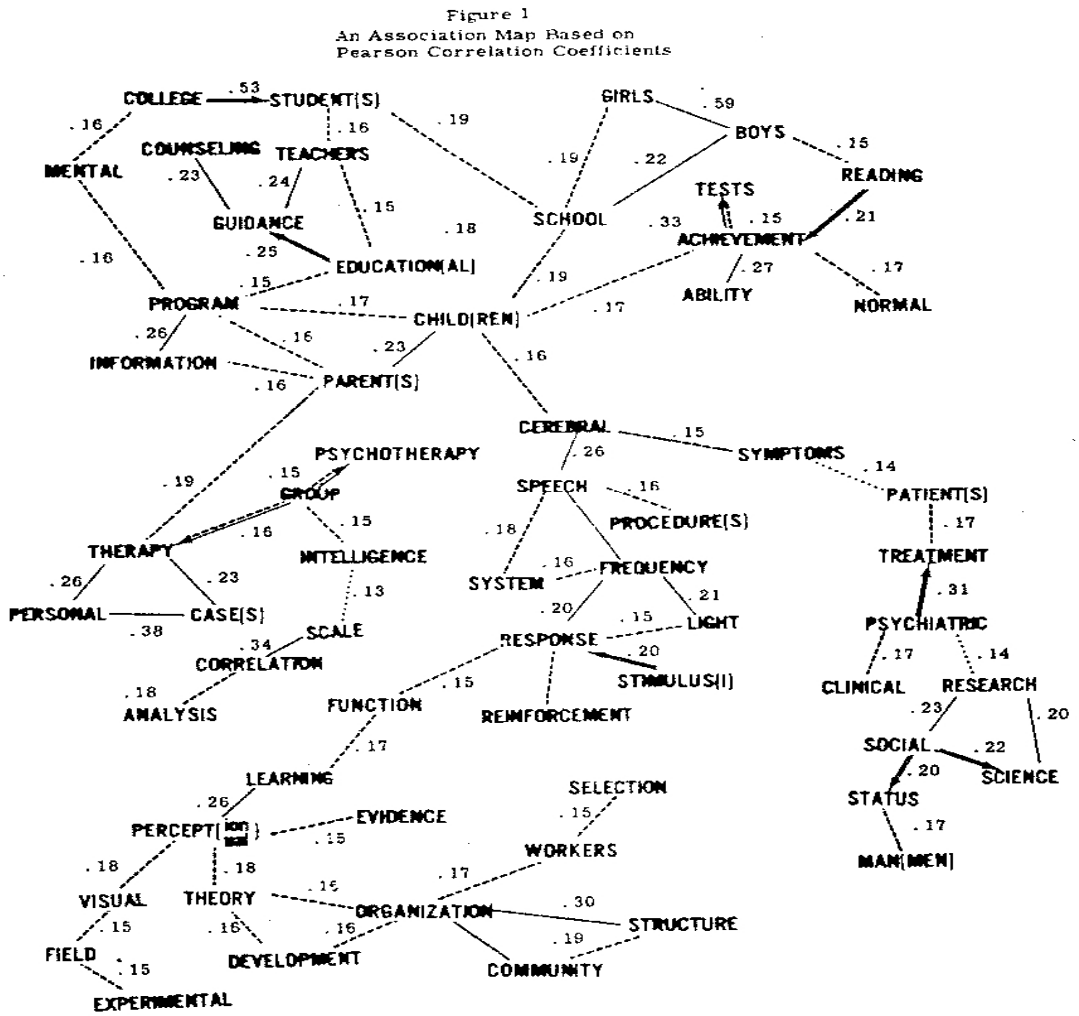



GRAFICO 2: Mapa de co-ocurrencia de palabras de Doyle aplicando Pearson.

# PARTE II

# TECNICAS

# DE

# PROCESAMIENTO

# TEXTUAL

# Capítulo 5

## Extracción y tratamiento de términos simples

Una visión común en la RI es ver al documento y a la interrogación del usuario como contenedores de palabras que serán comparados, de manera que, cuantas más palabras en común tengan, más relevante será el documento para esa búsqueda. Esta manera simplificada de mostrar la cuestión, no revela en su totalidad la verdadera complejidad de la automatización de un proceso que involucra al lenguaje. En este sentido, es que se cree importante destacar brevemente, algunos aspectos generales que las técnicas de indización deben considerar.

El primer punto es que no todas las palabras poseen el mismo nivel de significación para representar al documento. La teoría de la indización sugiere que algunas palabras conllevan más significado que otras. Por ejemplo, los sustantivos más que los adjetivos o los verbos, y todas ellas más que las preposiciones. Otro aspecto es que incluir todas las palabras de un texto acarrea ruido en la recuperación. El concepto de indización implica un vocabulario seleccionado, representar al documento solo con lo más significativo. Por último, un tercer aspecto, es que el lenguaje natural presenta muchas variaciones, y que al momento de buscar, es deseable expandir la búsqueda para incluirlas.

Existen dos grandes grupos de variaciones lingüísticas. El caso en el que dos expresiones distintas cargan con significados muy similares: sinonimia; o justamente lo opuesto, cuando dos formas iguales tienen distinto significado: polisemia. Otro caso más complejo es cuando una frase textualmente igual puede ser interpretada de manera diferente según el contexto. F. Kroon [44, p.27], citando a C. Jacquemin[1], sostiene que se puede distinguir tres tipos de variaciones de términos en RI:

**Variación sintáctica**: es cuando todas las palabras de la interrogación se encuentran presentes pero la estructura sintáctica es diferente. Por ejemplo: "*diseases of the lower urinary tract*" es una variación sintáctica de "*urinary tract disease*".

**Variación morfo-sintáctica**: además de presentar diferente sintaxis, presenta variación a nivel morfológico, pero las palabras siempre derivan de la misma raíz. Por ejemplo: "*translational or transcriptional inhibition*" es una variación morfo-sintáctica de "*translation inhibitor*".

[1] JAQUEMIN, C. *Spotting and discovering terms through Natural Language Processing*. Cambridge: MIT Press, 2001.

**Variación semántica**: además de poder presentar las variaciones anteriores, las palabras presentan diferencias de significado.

Intentando contemplar todos estos aspectos se han desarrollado muchas técnicas, algunas, verdaderamente sencillas; otras, más complejas, con fuerte aporte de la lingüística[2]. Se expone en este capítulo los procesos primarios que se aplican a los textos y una selección de aquellas técnicas que buscan dar solución al problema de las variaciones del lenguaje en la RI.

## 5.1 Identificación de unidades textuales

El análisis de los textos muestra diferentes niveles de estructura, organizados jerárquicamente de forma que una secuencia de frases se integra en estructuras mayores hasta conformar una unidad semántica global. J. Moreiro [13, p.30] sostiene que existe una **macroestructura general** que será el significado total de las **macroestructuras secundarias** que vinculan la información de cada parte entre las que se divide el texto. A su vez estas macroestructuras secundarias pueden dividirse en **macroestructuras parciales** de significado nodular a las que finalmente se podrá analizar en **microestructuras** de superficie, oraciones concretas de significado local. Esta clasificación moderna, proveniente del análisis del discurso, se completa para Moreiro con el concepto de **superestructura**, que alude a los modelos o formas textuales específicas de los distintos tipos de textos. Por ejemplo, la estructura tradicional de un trabajo científico en planteamiento del problema, materiales y métodos, discusión y comparación de resultados, conclusión.

Para la Documentación esto adquiere importancia, ya que al armar un SRI de carácter textual, el primer paso es la identificación de los textos reales o potenciales que integrarán la colección y formarán parte del repositorio. Dichos textos podrán presentar formas físicas diferentes, responder a tipologías documentales diversas e internamente presentar esquemas organizativos pertenecientes a diferentes superestructuras, lo que obliga, necesariamente, a una etapa de reconocimiento previo al diseño del sistema con la finalidad de determinar las unidades textuales que este procesará. Tal como se manifestó en el capítulo 1, la mayoría de los SRI utilizan una o varias bases de datos como estructura de almacenamiento y gestión de la información.

La unidad autónoma de información dentro de esta estructura es el **registro**. La unidad textual que se ha seleccionado estará, entonces, representada por un registro

[2] El Procesamiento del Lenguaje Natural (PLN) es el área de la lingüística dedicada específicamente a la investigación de los problemas que acarrea la automatización de procesos que involucren al lenguaje. Se comenzó a desarrollar fuertemente a partir de los años 70, como consecuencia de las investigaciones de Noam Chomsky sobre gramáticas

en la base de datos. El aspecto de la estructura interna, entonces, adquiere importancia, ya que, como producto de su análisis emanará el tamaño de la unidad textual a tratar. Como consecuencia, también, el tamaño del registro que el sistema deberá procesar. Esta decisión es crítica para una recuperación de información efectiva, ya sea porque se desea mostrar al usuario la porción de información precisa que él necesita o la optimización del funcionamiento de los algoritmos de búsqueda. Un registro muy pequeño brindará poco texto al algoritmo y este puede producir un resultado pobre; mientras que por el contrario, un registro muy extenso puede diluir la importancia del vocabulario y producir recuperaciones no relevantes.

D. Harman [45], reporta los siguientes datos sobre un experimento con colecciones de diferentes tamaños:

1) Un manual organizado en capítulos y secciones. Se tomó como registro cada párrafo.

2) Un código legal con secciones y subsecciones. Se tomó como registro las subsecciones.

3) Una colección de 40.000 casos de la corte. Se tomó como registro cada caso a texto completo.

| | Col.1 | Col.2 | Col.3 |
|---|---|---|---|
| Tamaño de la colección | 1.6 MB | 50 MB | 806 MB |
| Número de registros | 2653 | 6652 | 38304 |
| Prom. del número de términos por reg. incluidos duplicados | 96 | 1124 | 3264 |
| Número de términos únicos | 5123 | 25129 | 243470 |
| Promedio de "postings"[3] por término | 14 | 40 | 88 |

La mayoría de los experimentos sobre indización automática que han trabajado con artículos científicos han adoptado como unidad textual los datos de título, las palabras claves del autor y el resumen. El título, particularmente ha sido muy utilizado en Documentación ya que se supone que en el ámbito científico el autor trata de definir el tema sobre el que trata el trabajo en el título. Los primeros desarrollos dieron origen a los índices KWIC y KWOC. Salton sostiene que en muchas áreas, el agregar el texto completo en la recuperación agrega muy pocas mejoras [2, p. 71].

## 5.2 Identificación de unidades léxicas

generativas y funcionales. En el caso particular de la indización automática, las experimentaciones en extracción de términos se comparten con las áreas de la Terminología y la Traducción Automática.

[3] Posting: apuntador en el archivo invertido

Una vez que se han determinado las unidades textuales, el paso siguiente es la identificación de las palabras dentro de ellas, ya que un texto está compuesto además por espacios en blanco y signos de puntuación. Puede también contener guiones o números y palabras con mayúsculas. Todos estos elementos, que aportan valiosa información para muchos procesos automáticos en el tratamiento textual, deben aquí eliminarse o al menos se debe minimizar su intervención.

*Números*

Los números fuera de contexto no brindan demasiada información, por lo cual nunca serán buenos candidatos como términos de indización. Sin embargo no deben eliminarse los números que integran palabras. Por ejemplo, en el área de la óptica, la denominación del cristal fotorrefractivo “bi12sio20” contiene números y es deseable que se mantenga. En algunos contextos muy específicos, por ejemplo el DNI o el número de tarjeta de crédito podrían constituir buenos términos de indización, pero en la mayoría de los casos no son tenidos en cuenta. Para eliminarlos, se borran del texto de entrada todas las palabras que contienen secuencias de dígitos, menos las especificadas con una expresión determinada.

*Guiones*

Separar las palabras unidas con guión es conveniente para evitar inconsistencias, es decir que expresiones similares estén escritas con guión y sin guión. El problema aparece cuando el guión forma parte estructural de la palabra. Igualmente, como en el caso de los números, se eliminarán todos los guiones, salvo los específicamente indicados.

*Signos de puntuación*

En el caso de que un signo forme parte integral de la palabra, se opta por eliminarlo dado que se tomará la misma determinación con los términos de búsquedas utilizados por el usuario. Por ello, no habrá inconsistencias y la pérdida en la recuperación será mínima. Sin embargo, los signos de puntuación son utilizados por los analizadores gramaticales para encontrar agrupamientos de palabras, por lo cual, si el texto será analizado sintácticamente no serán eliminados en la primer instancia del tratamiento textual.

*Mayúsculas*

El caso de las mayúsculas y las minúsculas debe también analizarse en el contexto específico. Por lo general se convierten a un solo tipo elegido. En las interfaces de RI se observa con frecuencia que el término de búsqueda colocado en la interrogación es tratado por el sistema de manera indistinta (no-sensitiva) en cuanto a mayúsculas o minúsculas; salvo en aquellos casos en que explícitamente se permite señalar lo contrario. Esto puede acarrear problemas: transformar el término de búsqueda “AIDS” en “aids” no es lo más conveniente. [3, p.165] [44]

## 5.3 Eliminación de palabras no significativas

Como paso previo a la elección de los términos de indización es conveniente descartar los que con seguridad no serán buenos términos para la recuperación. Además, se considera que el tamaño de un texto se reduce un 40% después de eliminar las palabras no significativas. Una forma de hacerlo es confrontando cada palabra del texto con una lista previamente armada. Estas listas suelen llamarse “antidiccionario” o “lista de palabras vacías”. Están compuestas por los artículos, adverbios, pronombres, preposiciones, conjunciones, exclamaciones. Existen diversas listas para los distintos idiomas. Otro método es producir un listado de frecuencias de palabras del texto a indizar y luego examinar las palabras de frecuencia muy alta. Esto es lo que realizó el NIST (National Institute of Standards and Technology) con la colección del Wall Street Journal, que es una de las usadas en los experimentos de evaluación de los TREC. Las palabras removidas de las 29 más frecuentes fueron "a", "at", "from" y "to". En general los sistemas comerciales como ORBIT inclusive el mismo MEDLARS usan muy pocas palabras no significativas [45].

## 5.4 Conflación

La conflación (fusión) de términos en Recuperación de Información, es la reducción de la variedad lingüística de los documentos por medio de la agrupación de las ocurrencias textuales que se refieren a conceptos similares o idénticos [46]. Las técnicas más comunes de conflación son la **truncación**, el **stemming**[4], la **lematización** y la aplicación de **diccionarios de sinónimos**.

[4] Se ha decidido mantener el término en inglés ya que no existe una equivalencia exacta en castellano. La traducción más aproximada sería “reducción a la raíz”, pero en rigor esta es una tarea netamente lingüística. El stemming se basan en esa idea, pero es una “reducción a la raíz” hecha por las computadoras.

La truncación es una técnica no-lingüística que se basa únicamente en la comparación de cadenas de caracteres. Es muy fácil de implementar y la mayoría de los SRI la aplican. La truncación a la derecha es la más usual y permite realizar una búsqueda conjunta de todas las palabras que comparten un mismo comienzo.

La técnica de stemming trabaja sobre el supuesto de que todas las palabras con la misma raíz lingüística están conceptualmente relacionadas. Por ejemplo, en el idioma inglés, las palabras *magnesia, magnesian, magnet, magnetic,* serán fundidas en la raíz *magnes.* La cadena de caracteres resultante de aplicar este proceso se la denomina raíz y, aunque no necesariamente es igual a la raíz lingüística, como mínimo debe servir para desambiguar el término.

La lematización corresponde a un nivel de análisis morfológico-léxico del texto que busca reducir las variaciones enviando las palabras hacia su forma canónica o entrada léxica en un diccionario. Por ejemplo, los verbos en forma de infinitivo, los sustantivos en singular y los adjetivos en masculino singular [13, p.123]. Este tipo de técnicas son más costosas y trabajan utilizando herramientas de Procesamiento del Lenguaje Natural (PLN) tales como los “parsers lingüísticos”.

La detección de sinónimos se usa para resolver el problema de que dos expresiones lingüísticas distintas cargan con significados muy similares. En este caso se supone que para el usuario, la información que contiene uno u otro término será relevante.

Para algunos idiomas, los problemas que acarrean las variaciones lingüísticas en la RI son muy significativos. Para el hebreo se ha reportado que solo entre un 2% y un 10% se recupera si no se aplican técnicas que traten este problema. El inglés es un idioma con pocas variantes, típicamente solo posee 3 ó 4 por palabra [46].

### 5.4.1. Stemming

Cuando se realiza la extracción de palabras de un texto se obtiene una gran cantidad de entradas con formas verbales conjugadas y variantes de concordancia. Logrando la reducción morfológica de todas estas variantes se busca que el usuario recupere tanto los textos que contienen sus términos de búsqueda, como aquellos que contienen las formas derivadas de esos términos. Por ejemplo, en idioma inglés, *analysis, analyzing, analyzer, analysing* puede reducirse a la forma *"analy"* que se considera su raíz. Los plurales, los gerundios y las formas de los verbos en pasado son los casos más comunes de palabras susceptibles de aplicar esta técnica. Durante

el proceso de reducción siempre existirá un porcentaje de error, pero este es lo suficientemente bajo como para no afectar a la efectividad en la recuperación[5].

Los errores más frecuentes tienen que ver con la "sobre-reducción" y la "sub-reducción". Por ejemplo reducir *centennial*, *century*, *centre* y *center* en la raiz *cent*. Por el contrario, reducir *acquire*, *acquiring* y *acquired* en la raíz *acquir* y reducir *acquisition* en *acquis*. Otro caso común de error se da cuando se desea recuperar información en **colecciones** multidisciplinares. Por ejemplo, el término "*living*" en el ámbito de la biología puede reducirse a "*live",* pero no seria deseable en el área de arquitectura [47].

Clasificar a este tipo de algoritmos no es tarea sencilla. Los diferentes autores no acuerdan al respecto, y es evidente que es imposible establecer clasificaciones taxativas. En sentido general los autores establecen que existen **algoritmos lingüísticos** y **algoritmos no-lingüísticos**. Resulta difícil suponer que se pueda construir una técnica de este tipo sin tener en cuenta aspectos de la lengua, por lo cual, esto debe verse más como dos enfoques diferentes al producir estrategias de construcción. Dentro de los primeros, los hay que

- únicamente trabajan sobre la **inflexión**.
- únicamente trabajan sobre la **derivación**[6].
- utilizan formas combinadas de los anteriores.

Dentro de los segundos, R. Baeza-Yates [3, p.169], cita a W. Frakes[7] quien sostiene que existen:

- algoritmos que remueven afijos
- algoritmos que contrastan contra una tabla de raíces
- algoritmos basados en el sucesor variante

[5] C. van Rijsbergen [6, p.12] menciona un error del 5 %.

[6] La Morfología es el área de la lingüística que estudia la estructura internas de las palabras. Usualmente se divide en dos subclases en
- inflexional
- derivacional

La morfología inflexional describe los cambios predecibles de las palabras producidos por la sintaxis: los plurales, las formas posesivas y los tiempos verbales son las más comunes. Estos cambios no tienen efecto sobre la palabra como parte del habla (part-of-speech), el sustantivo sigue siendo sustantivo después de la pluralización. En contraste, la morfología derivativa puede o no afectar a la palabra como parte del habla y puede o no afectar su significado. Por ejemplo, en inglés los sufijos "-ize" y "-ship": *member / membership.*

[7] Frakes, W.B. y Baeza-Yates, R. *Data structures and algorithms*. Englewood Cliffs, NJ: Prentice Hall, 1992.

- algoritmos basados en N-grams

El método más común es el de remover los afijos. En el idioma inglés, principalmente se remueven los sufijos que son los que provocan la mayoría de las variantes. Uno de los más simples es el "S-stemmer" que se usa para eliminar los plurales. Trabaja removiendo solo terminaciones simples: "ies", "es" y "s", salvo algunas excepciones que se consignan en una lista.

*sses* ⟶ *ss;*
*ies* ⟶ *i;*
*ss* ⟶ *ss;*
*s* ⟶ Ø;

Así, cuando se encuentra la palabra *stresses,* la cadena *sses* es reemplazada por *ss* para generar el singular *stress* en lugar de *stresse* que sería el resultado de aplicar la última de estas reglas. Esto es solo un ejemplo de la forma en que operan estos algoritmos.

Un avance hacia la eliminación de otro tipo de sufijos lo produjo el algoritmo desarrollado por J. Lovins[8] en 1968. Luego, basándose en él, M. Porter [48] desarrolló en 1980 el que sería el más implementado en la mayoría de los SRI de las dos décadas siguientes. Este algoritmo utiliza una tabla de sufijos y se contrastan las palabras contra ella para proceder a su eliminación. Como suele presentar problemas, se brindan reglas de contexto. Por ejemplo es deseable eliminar UAL de FACTUAL, pero no de EQUAL. Una regla de contexto, por ejemplo, establecería que se elimine UAL siempre que no lo preceda una Q. Cuando la reducción de los sufijos no actúa bien, lo que suele hacerse es construir una lista de terminación de raíces equivalentes. Dos raíces que serán tomadas como equivalentes, deberán ser iguales solo en sus terminaciones que estarán en la lista de equivalencias. Por ejemplo las raíces ABSORB- y ABSORPT- serán fundidas, porque las terminaciones B y PT figuran como equivalentes cuando las precede la misma raíz. El método de contrastar los términos contra una tabla de raíces preexistente es el método más fácil de implementar en términos computacionales, pero difícil de poner en práctica porque requiere tener almacenadas todas las raíces de una lengua determinada. Es más costoso comparado con la técnica de eliminación de sufijos.

[8] LOVINS, J.B. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*. 1968, v.11, nro, 1-2, p.22-31.

D. Harman [45] brinda los siguientes resultados al aplicar los algoritmos S, Porter y Lovins a una búsqueda del test Cranfield: *“panels subjected to aerodynamic heating”*

| S | PORTER | LOVINS |
|---|---|---|
| panel | panel | Panel |
| panels | panels | Panels |
| subjected | subjected | Subjected |
| | subjective | Subjective |
| | subjects | Subjects |
| aerodynamic | aerodynamic | Aerodynamic |
| aerodynamics | aerodynamics | Aerodynamics |
| | aerodynamically | Aerodynamically |
| | | Aerodynamicist |
| heating | heating | Heating |
| | heated | Heated |
| | | Heat |
| | | Heats |
| | | Heater |

El método del sucesor variante se basa en determinar los morfemas del contorno de acuerdo a la lingüística estructural. Es bastante más difícil de implementar.

El método de los N-gramas se debe pensar como una ventana de n-caracteres de tamaño que se va desplazando a través del texto del documento y de la consulta. Así, si n=3, se obtienen una serie de trigramas que serán sometidos al mismo tratamiento que se le dá a los términos en el modelo de recuperación vectorial. Se calculan sus frecuencias, sus pesos y la similaridad con los trigramas de la búsqueda. Por ejemplo, la palabra *biblioteca*, produciría los siguientes trigramas: _bi, *bib, ibl, bli, lio, iot, ote, tec, eca, ca*_. De esta manera, si se tienen las palabras *bibliotecas, bibliotecarios, bibliotecología, biblioteconomía*, gran parte de los n-gramas serán iguales, lo que en un cálculo de similitud dará mucha cercanía entre ellos.

#### 5.4.1.1. Evaluación

Los experimentos llevados a cabo para evaluar los diferentes tipos de algoritmos han sido numerosos, principalmente en la década del noventa. Solo se exponen aquí algunas de las conclusiones más significativas a las que ellos han arribado.

En 1990, W. Frakes [49] realizó un par de experimentos sobre la base de datos del Psychological Abstract. Utilizó el motor de búsqueda de DIALOG, el cual aplica la técnica de reducción a la raíz basada en el algoritmo de Lovins. Como resultado de su evaluación determinó que:

- Los usuarios, al truncar términos durante el proceso de búsqueda, lo hacen a nivel de la raíz de forma natural. Si no, al menos, la desviación es muy pequeña.
- Esta desviación no afecta significativamente la efectividad en la recuperación de información.
- Para la recuperación, no existe diferencia significativa entre la conflación realizada manualmente y la realizada de manera automática.

En 1991, D. Harman [50] realizó un estudio comparativo del desempeño de tres algoritmos: S-stemmer, Lovins y Porter. Como resultado de su estudio concluyó que para el idioma inglés, ninguno de los tres producía mejoras en la efectividad de la RI. Después de una detallada evaluación, obtuvo que el número de interrogaciones que se beneficiaban aplicando los algoritmos era igual al que se deterioraba.

En 1992, M. Popovic y P. Willett [51] evaluaron el desempeño del algoritmo de Porter en una colección textual en idioma esloveno. El experimento arrojó muy buenos resultados en cuanto a la evaluación de la precisión. Luego procedieron a realizar el mismo experimento sobre la misma colección traducida al inglés. Los resultados obtenidos fueron similares a los de Harman, por lo cual, los autores concluyen que la eficacia de estos algoritmos depende de la complejidad morfológica del lenguaje sobre el cual se aplican.

En 1993, R. Krovetz [47] evaluó cuatro algoritmos de tipo lingüístico y los comparó con el algoritmo de Porter. Obtuvo resultados muy favorables en cuanto a las medidas de exhaustividad y precisión respecto a no aplicar la técnica de reducción a la raíz. Los mejores resultados se daban cuando los documentos eran cortos. Estos experimentos fueron realizados sobre cuatro colecciones de evaluación en inglés, una de ellas la CACM (ACM) que fue la misma que utilizó Harman. Sin embargo, hay que destacar que ambos experimentos utilizando la misma colección y aplicando las mismas medidas, arrojaron resultados muy diferentes al evaluar el algoritmo de Porter [48].

En 1996, D. Hull [52] argumentó que las tres medidas que habitualmente se aplicaban para la evaluación de estos algoritmos no eran suficientes. A la precisión promedio, exhaustividad promedio y precisión promedio en los 11 puntos de exhastividad (conocida como [APR11] *Average Precision Recall), a*gregó la de promedio de precisión en los 5-15 documentos examinados AP[5-15] y promedio de exhaustividad en los los 50-150 documentos examinados AR[50-150]. Examinó la colección TREC con 5 algoritmos de reducción a la raíz diferentes. Concluyó que su aplicación siempre es beneficiosa, menos en el caso de interrogaciones muy extensas que arrojan muy bajo nivel de exhaustividad. Mostró además que existen diferencias a favor de los algoritmos lingüísticos respecto a los del tipo de Porter y Lovins.

En 1996, W. Kraaij y R. Pohlmann [53] experimentaron con una versión del algoritmo de Porter para el idioma alemán. Aplicaron una herramienta de recuperación basada en el modelo del espacio vectorial que incorporaba la técnica de expansión de la interrogación[9]. Ponderaron de manera diferente a los términos puros de los que se agregaban por variaciones. Evaluaron dos algoritmos de reducción a la raíz, uno inflexional y otro derivativo, usando la base de datos léxica CELEX. Expandieron las búsquedas con las variantes lingüísticas de los términos que se encontraban en esa misma colección de prueba y encontraron que sobre 50.000 formas de palabras diferentes, el 40% no estaba incluido en la base CELEX. Se examinaron 2500 al azar y se encontró que

46% Nombres propios
37% frases nominales (Nominal compounds)
10% errores ortográficos
3% otras lenguas
3% variantes morfológicas que no estaba en CELEX
1% Raíces que no estaban en CELEX

Las principales conclusiones a las que arribaron fueron:

- La técnica de reducción a la raíz mejora la exhaustividad a costa de la precisión
- El desempeño del algoritmo de Porter y los lingüísticos basados en CELEX no mostraban diferencias.

[9] En inglés se denomina “Query expansion” y significa que se amplía la interrogación con términos relacionados que aparecen en un diccionario. Por ejemplo, en el modelo del espacio vectorial, esto permite que las frecuencias de aparición de las variantes morfológicas de un mismo término sean agrupadas para el cálculo de la similitud.

- La técnica de reducción a la raíz aplicada durante la expansión de la búsqueda, brinda el mismo nivel de efectividad que si se aplica durante la indización.
- Aplicar la técnica de manera selectiva (solo inflexional) arroja mejores resultados que aplicarla de manera completa (inflexional y derivativo).

En 1998, M. Fuller y J. Zobel [54] realizaron un experimento utilizando un vocabulario de 363.553 términos distintos pertenecientes a la colección de datos de TREC 2 y 1.093 términos correspondientes a 75 interrogaciones diferentes también pertenecientes al test de TREC. Le aplicaron al vocabulario los diferentes algoritmos: S, Porter, Lovins y un diccionario de raíces propio, con la finalidad de determinar el conjunto de conflaciones para cada término de búsqueda. Así obtuvieron un promedio de 13,5 conflaciones por término. Luego revisaron manualmente la calidad de lo obtenido, y observaron que el promedio de conflaciones correctas era menor: 7,9. Considerando ese valor como el deseado, los resultados obtenidos se resumen en la tabla siguiente:

| | Intentos | Correctos | Perdidos |
|---|---|---|---|
| Sin stemming | 1,0 | 1,0 | 6,9 |
| Stemming perfecto | 7,9 | 7,9 | 0,0 |
| S | 1,7 | 1,7 | 6,1 |
| Diccionario | 4,1 | 3,8 | 4,1 |
| Porter | 6,3 | 5,2 | 2,6 |
| Lovins | 11,1 | 5,9 | 2,0 |

En esta tabla se observan los valores promedio de conflaciones intentadas por cada término y cuántas fueron correctas. Así se observa que el algoritmo S, que siempre se lo consideró como uno de los más confiables, solo encuentra un 0,7 de términos adicionales, es decir un 22% de todos los posibles. En el otro extremo, el algoritmo de Lovins parece más agresivo, pero comparado con el de Porter, con 4,8 conflaciones más por término de búsqueda, solo 0,7 términos adicionales son correctos. Lovins posee una exactitud del 53%, mientras que el de Porter refleja una precisión del 83%. El diccionario de raíces es el más exacto, 93%, sin embargo pierde varios términos que son encontrados por los algoritmos de Porter y de Lovins. Estos tres métodos encuentran 48%, 66% y 75% de conflaciones correctas, respectivamente.

**Comentario [C1]:** 1,7 / 1
**Comentario [C2]:** 1,7 / 7,9
**Comentario [C3]:** 11,1 - 6,3
**Comentario [C4]:** 5,9 - 5,2
**Comentario [C5]:** 5,9 / 11,1
**Comentario [C6]:** 5,2 / 6,3
**Comentario [C7]:** 3,8 / 4,1
**Comentario [C8]:** 3, 8 / 7,9
**Comentario [C9]:** 5,2 / 7,9
**Comentario [C10]:** 5,9 / 7,9

# Capítulo 6

# La ponderación de términos simples

La ponderación es una técnica que se aplica en los SRI con la finalidad de mejorar la calidad de la recuperación. Existen diversas variantes y los investigadores y desarrolladores las han aplicado de diferentes maneras dentro de los algoritmos de búsqueda. El resultado más conocido de su aplicación es la obtención de salidas ordenadas, pero también, tal como se ha expuesto en el capítulo 2, puede cumplir un rol importante dentro del propio corazón del sistema en el momento de la construcción de la representación. Una vez que el vocabulario es seleccionado de acuerdo a las técnicas expuestas en el capítulo anterior, se realiza una "valoración" de cuales son los mejores términos para representar al documento.

En un primer análisis se puede sostener que existen tres tipos diferentes de ponderación:

1. la ponderación que realiza el usuario cuando desea destacar un término de búsqueda sobre otro en su interrogación.
2. la ponderación de los términos dentro de un mismo documento.
3. la ponderación orientada a los sistemas, que refleja el comportamiento de los términos en la totalidad de la colección.

El uso en un SRI de la ponderación señalada en 1, contribuye a que el usuario logre representar de mejor manera su necesidad de información. Las ponderaciones 2 y 3, usadas generalmente de manera conjunta, contribuyen a una mejor representación del documento en el contexto de la colección.

## 6.1 Principios

Algunas de las ideas que sirven como antecedentes a ésta técnica ya han sido introducidas en los capítulos previos, específicamente lo referido a tomar las palabras y su frecuencia como indicadores de significado. Sin embargo, para su completa comprensión, también es necesario revisar un par de conceptos propios de nuestra disciplina: exhaustividad y especificidad aplicados a la RI.

### 6.1.1 Exhaustividad y especificidad

Las nociones de **exhaustividad en la indización** y **especificidad de los términos de indización** son conceptos que hacen referencia, en el primer caso, a cierta propiedad de la descripción de contenido, y en el segundo, a cierta propiedad que posee un término de indización en particular. Si se analiza la indización de un documento, la **exhaustividad** está dada por la cobertura que hacen los términos asignados de los diferentes temas que trata el documento. Mientras que la **especificidad** de uno de esos términos es el nivel de detalle con el cual representa al concepto [6, p.14].

La relación que presenta el concepto de exhaustividad con la ponderación de términos, parte de interpretar que la exhaustividad aumenta si se le asignan más términos de indización al documento. Cuando el número de términos del vocabulario de indización es constante, la probabilidad de que el documento sea recuperado crece. Por ello se sostiene que el aumento de la exhaustividad mejora la performance de tipo “recall” del sistema de recuperación. Ahora bien, cuando la descripción de contenido es más grande, habrá términos más usados frecuentemente. Esto es inevitable con vocabularios controlados (tamaño constante), pero también es aplicable a la extracción de palabras del texto, especialmente si se aplica stemming (debe recordarse que esta técnica reduce la variabilidad del lenguaje natural). Entonces, el crecimiento de las palabras del vocabulario no crece pareja con el crecimiento del número de documentos indexados. La extracción de más palabras del documento hará que aumente la frecuencia de la palabra más que generar palabras nuevas. Cuanto más exhaustiva sea la indización, más términos se usarán y su frecuencia de uso aumentará. Esto provoca que el término se transforme en menos efectivo para la recuperación dado que no discrimina. Hace que los documentos no se puedan distinguir entre sí.

La especificidad es una propiedad de un término de indización. Según Spark Jones [55] es una característica semántica de los términos de indización, donde un término es más o menos específico si su significado es más o menos detallado o preciso. Cuando se construye un vocabulario de indización, se toman diversas decisiones acerca del poder de discriminación de cada término de acuerdo con su propiedad descriptiva. Por ejemplo: la decisión de incluir “infusiones” o de incluir “té”, “café” o “cacao”. Estas decisiones generalmente se toman pensando en la necesidad de distinguir entre los documentos en el momento de la recuperación. Spark Jones sostiene:

«.... *given level of indexing exhaustivity is believed to be sufficient to represent the content of individual documents adequately, and distinguish one document from another»*

El término más general: “infusiones”, será asignado a muchos más documentos que si se tratara de alguno de los términos específicos –siempre que se construya el vocabulario eligiendo exclusivamente una de las dos alternativas. Un aumento en la especificidad de los términos utilizados aumenta la performance de “precisión” del sistema de recuperación y desde el punto de vista estadístico, cuanto más precisos sean los términos, menos frecuentes serán en la colección.

Pensando en una posible manera de medir o representar algorítmicamente ambos conceptos, Spark Jones redefine a la exhaustividad de la descripción de un documento como el número de términos que contiene, y la especificidad de un término como el número de documentos que lo contienen. Esta visión cuantitativa del problema, es la que permitió el posterior desarrollo de las principales funciones matemáticas de ponderación en las cuales se contempla la capacidad de discriminación de los términos a partir de su frecuencia en la colección.

### 6.1.2 Consideraciones sobre la frecuencia

- A mayor frecuencia de aparición de un término en un documento, mayor será su peso, dado que se considera que el documento trata sobre los conceptos que explicitan los términos más frecuentes.
- Sin embargo, existe una cantidad de términos que son muy frecuentes y poco útiles. Estos deben ser descartados del vocabulario, por ejemplo con una lista de palabras no significativas.
- Lo mismo sucede con los términos que son muy poco frecuentes, ya que existe una muy baja probabilidad de que un usuario busque por alguno de ellos. También son eliminados.
- Al aumentar el tamaño de la colección aumenta la frecuencia de aparición de cada término. El término sigue siendo relevante aunque la frecuencia sea alta si la colección es grande.
- Cuando más largo es el documento hay más probabilidad de repetir palabras. Si en dos documentos, un mismo término aparece el mismo número de veces, se considerará que el peso del término en el documento más corto es mayor.
- La eficacia de una palabra como término de indización depende de la colección. Interesan términos que clasifiquen la colección, diferenciando documentos relevantes de los irrelevantes para una búsqueda dada. Por ejemplo *digital* será

mal discriminador en una colección de *Informática*, sin embargo será bueno en una colección de *Filosofía.*

### 6.1.3 Pasos para el cálculo de los pesos:

Una vez que se tiene el texto tratado según los pasos del capítulo anterior, al menos en lo que respecta a la selección del vocabulario, se procede a:

1) Calcular la frecuencia de aparición de cada palabra en cada documento
2) Calcular la aparición de cada palabra en toda la colección (se suman las frecuencias del punto 1).
3) Calcular el número de términos en cada documento.
4) Calcular el número de documentos en los que aparece cada término.
5) Construir una matriz de asociación de términos y documentos. Esta matriz contiene tantas columnas como documentos tiene la colección y tantas filas como palabras tiene el vocabulario. Por ejemplo:

| | D1 | D2 | D3 | … | $D_j$ | … | $D_n$ | Frecuencia Total (**$tf_i$**) | Número documentos (**N**) |
|---|---|---|---|---|---|---|---|---|---|
| $T_1$ | 0 | 0 | 1 | … | 1 | 1 | 1 | 4 | 4 |
| $T_2$ | 0 | 2 | 1 | … | 0 | 0 | 1 | 4 | 3 |
| … | … | … | … | … | … | … | … | | |
| $T_i$ | 1 | 0 | 0 | … | 0 | 1 | 0 | 2 | 2 |
| … | … | … | … | … | … | … | … | | |
| $T_m$ | 2 | 0 | 0 | … | 0 | 0 | 0 | 2 | 1 |
| $NT_j$ | 2 | 1 | 2 | … | 1 | 2 | 2 | | |

- $NT_j$ = Número de términos en **$D_j$**
- Cada celda contiene el valor absoluto de la frecuencia del término *i* en el documento *j* y se denota como **$f_{ij}$**.
- La Frecuencia total es la cantidad de veces que aparece el término en la colección y se denota por **$tf_i$**.
- El número de documentos en que aparece cada término se denota como **$fD_j$**.

6) Luego, se aplica el concepto de Luhn para eliminar las palabras muy frecuentes y las muy poco frecuentes. R. Peña [15, p.255] sostiene que en un sistema se puede optar por eliminar un porcentaje de los términos que

aparecen en muchos documentos, esto es, que posean un valor de $\mathbf{fD_j}$ muy grande. O también eliminar los que tienen una frecuencia total en la colección mayor que un umbral determinado $\mathbf{tf_i > maxft}$. De manera similar se puede establecer un umbral de frecuencia mínima necesaria para formar parte del vocabulario. Por lo general esto varía de acuerdo al tamaño de la base de datos.

## 6.2 Tipos de ponderación

Existen distintas maneras de calcular el peso de un término en un documento teniendo en cuenta la colección en la cual ese documento está inmerso. Se detallan aquí únicamente las que se consideran aproximaciones diferentes, realizadas por autores cuyas investigaciones han resultado muy significativas en este campo y que luego los experimentos en RI han aplicado en combinaciones diversas.

### 6.2.1 Ponderación basada en la relación término/documento

- **Ponderación binaria**

  El peso de un término tendrá valor **1** si el término está presente ó **0** si el término está ausente en el documento.

- **Ponderación por frecuencia absoluta**

  El peso de un término en un documento dado tendrá el valor de su frecuencia.

### 6.2.2 Ponderación basada en la relación término/documento/colección

- **Ponderación basada en la frecuencia absoluta del término**

  Conocida como **tf** es una forma de ponderación muy sencilla. Considera que los términos que son frecuentemente mencionados en documentos individuales pueden ser más útiles que otros a los fines de la recuperación. Brinda muy baja "performance" al SRI, dado que cuando un término tiene frecuencia alta y no se encuentra concentrado en algunos documentos particulares de la colección, todos los documentos tienden a ser recuperados y esto afecta las medidas de precisión del sistema. Salton y Yang [56] la definen como:

$$tf_i = \sum_{j=1}^{n} f_{ij}$$

- **Ponderación basada en la frecuencia del término normalizada**

  Esta función normaliza la frecuencia del término para que solo tome valores entre 0.5 y 1

$$P_{ij} = 0.5 + 0.5\frac{tf}{\max tf}$$

- **Ponderación basada en el tamaño del documento**

  En esta ponderación el peso de cada término se calcula como el cociente entre la frecuencia absoluta del término **$tf_{ij}$** y el tamaño del documento **$NT_j$**, es decir la cantidad de términos que posee:

$$P_{ij} = \frac{tf_{ij}}{NT_j}$$

- **Ponderación basada en la frecuencia inversa (I)**

  Fue propuesta por K. Sparck Jones [55] buscando introducir un factor de dependencia de la colección y ponderar mejor la concentración de términos en unos pocos documentos. Asume que la importancia de un término es proporcional a la frecuencia que dicho término tiene en cada documento e inversamente proporcional al número total de documentos en que dicho término es asignado. Se la denomina **idf** y puede expresarse como:

$$P_{ij} = \log_2 \frac{n}{N} + 1$$

  Donde **n** es el número total de documentos de la colección y ***N*** es el número de documentos que contienen el término. Cuando ***N*** disminuye **$idf_i$** aumenta. Como ***N*** es un número comprendido entre ***n*** y 1 **(*n* >*N* >1)**, **$idf_i$** toma el valor mínimo 1, aumentando hasta un valor constante **(1 < idf < $\log_2(n)$+1)**. Por propiedades del logaritmo esta función puede expresarse como:

$$P_{ij} = \log_2(n) - \log_2(N) + 1$$

Este tipo de ponderación sugiere que deberán tomarse como descriptores aquellos términos con frecuencias superiores a la media. Una vez que el término ha sido seleccionado para integrar el vocabulario será utilizado como descriptor en todos los documentos que lo contengan.

- **Ponderación basada en la frecuencia inversa (II)**

  Conocida como **tf * idf**, esta función se basa en el concepto de que los mejores términos para identificar son aquellos que permiten distinguir ciertos documentos individuales dentro del resto de la colección. Esto implica que los términos buenos tendrán una frecuencia alta en el documento, pero baja en la colección. Un ajuste es propuesto por Salton [2, p.63] al incorporar las dos ponderaciones expuestas en los puntos anteriores en una sola función:

$$P_{ij} = tf_{ij} * \left[\log_2(n) - \log_2(N) + 1\right]$$

### 6.2.3 Ponderación basada en el factor normalizado de vectores

Esta función incorpora la longitud del vector de términos de cada documento como factor a considerar. En muchas situaciones, documentos cortos suelen quedar representados por vectores cortos y documentos más extensos por vectores más largos. Cuando un número más grande de términos representa al documento, este tendrá más posibilidades de ser recuperado ya que la posibilidad de que alguno de sus elementos se equipare con algún elemento de la expresión de búsqueda es mayor [57]. Se incorpora entonces, en la función del cálculo del peso, un factor que iguala la longitud de los vectores documentales, dividiendo el peso del término por la suma de los pesos de todos los términos que representan a ese documento:

$$P_{ij} = \frac{tf_{ij} * (\log_2 \frac{n}{N} + 1)}{\sum_{vector} (tf_{ij} * (\log_2 \frac{n}{N} + 1)^2}$$

### 6.2.4 Ponderación basada en el poder de discriminación de un término

En la exposición del modelo del espacio vectorial realizada en el capítulo 2 de este mismo trabajo, se introdujo el tema del cálculo de la **similitud** entre vectores que

representan documentos y vectores que representan búsquedas. Basándose en ese mismo concepto, Salton y Yang [56] proponen la idea de que un cálculo de similitud entre cada par posible de vectores documentales de una colección servirá para ponderar mejor a cada término en lo que respecta a su "**poder de discriminación**"; es decir, la capacidad de cada término dentro de una colección dada para distinguir un documento de otro. Si la similitud entre el vector documental $\boldsymbol{D_j}$ y el vector documental $\boldsymbol{D_k}$ se expresa como la sumatoria de los productos de los pesos:

$$Q_{jk} = \sum_{i=1}^{n} P_{ij} * P_{ik}$$

Se puede obtener un valor promedio de ese valor de similitud:

$$\overline{Q} = \frac{2}{n(n-1)} \sum_{i-1}^{n-1} \sum_{j=i+1}^{n} Q_{ij}$$

Este promedio en realidad es un reflejo de la densidad de ese espacio documental concreto, es decir de los documentos de esa base de datos indizados con ese conjunto de términos particular. Para mantener normalizado el valor de la similitud entre 0 y 1, se introduce en la función un valor constante. Salton [2, p.66] propone que ese valor sea 1/*n*(*n*-1) y Peña [15, p.259 ] propone que sea 2/*n*(*n*-1). Cuando todos los *n* documentos son idénticos, el valor de similitud es igual a 1 y la similitud promedio alcanza el valor máximo.

La densidad del espacio documental puede ser calculada de manera más eficiente si se construye un documento **centroide** en el cual los términos asumen un valor de frecuencia promedio. Es decir, la frecuencia del término *T* es definida por:

$$\overline{f_i} = \frac{1}{n} \sum_{i=1}^{n} f_{ij}$$

Entonces, se calcula la densidad como la suma de las similitudes de cada documento con el centroide, donde la densidad variará entre 0 y *n*. Un valor más alto de densidad, indica un nivel más alto de compactación en el espacio documental.

La contribución de un término cualquiera *m* a la densidad del espacio documental, es decir su valor de discriminación, puede ser calculada mediante la función $Q_m - Q$, donde $Q_m$ es la compactación del espacio documental con el término *m* eliminado. Si el término *m* es un buen discriminador, entonces $Q_m$ será > que *Q,* esto es, el espacio

documental después de remover el término se volvió más compacto. Si el término *m* es un término amplio, de alta frecuencia, con una distribución de frecuencias pareja, esto es, que puede aparecer en muchas descripciones documentales, entonces si lo eliminamos, se reducirá la similitud promedio para cada par documental. Luego de calcular todos los valores de discriminación para cada término se los podrá ordenar de manera descendente según su valor de discriminación. Luego, Salton [2, p.67] propone clasificar a los términos en tres grupos:

1. Los que se considerarán valores positivos, y que al usarlos en la indización harán descender la densidad del espacio documental.
2. los que se considerarán valores medios, cercanos a 0, cuya inclusión o eliminación no alteran el valor de similitud entre los documentos.
3. los que se considerarán valores negativos, cuya utilización hace que los documentos sean más similares unos de otros.

Según Moreiro [13, p.107] se considera que los términos que aparecen en el 80% de los documentos de una colección no son útiles para la recuperación de información.

### 6.2.5 Ponderación basada en la probabilidad de relevancia

En el capítulo 2 se expuso de manera general el modelo de SRI basado en cálculo de probabilidades. Este modelo incorpora el concepto de "ponderación por relevancia". La idea original es de S.E. Robertson [58] quien propuso utilizar la información disponible de como se distribuyen los términos en los **documentos relevantes**. En los modelos de ponderación anteriores, un término cualquiera tiene el mismo peso para búsquedas diferentes, en cambio, cuando los documentos relevantes son tenidos en cuenta, un mismo término puede pesar diferente según las distintas búsquedas.

S.E. Robertson y K. Sparck Jones [59] proponen una función básica de ponderación que es:

$$p_{ij} = \log \frac{\left(\frac{r}{R}\right)}{\left(\frac{n}{N}\right)}$$

donde $r$ es el número de documentos relevantes indizados con el término y $R$ es el número de documentos relevantes para una búsqueda dada. La cuestión aquí es cómo determinar cúales son los documentos relevantes. Según Robertson y Sparck

Jones, Barkla propone que un servicio de DSI puede proveer la información por el feed-back constante que se establece y Miller propone que sea estimado por el usuario. Este tipo de ponderación presenta muy buena “performance” cuando se aplica a colecciones de prueba ya que la información sobre la relevancia perfecta está disponible. Un ejemplo de una colección de prueba es la colección Cranfield empleada en la parte experimental de este trabajo.

Básicamente se asume una descripción binaria de los documentos (valores 0 o 1 en la matriz término/documento), luego se asume que existe un conjunto de juicios de relevancia para cada interrogación. Los documentos serán juzgados como relevantes respecto a la interrogación en relación con alguna necesidad particular, es decir que el juicio de relevancia es específico para un usuario dado. Otro usuario con la misma interrogación en términos verbales, pero con un juicio de relevancia diferente, necesitará una estrategia diferente. Es decir, siempre se considera que las interrogaciones reflejan necesidades individuales.

Entonces, dado una colección, un término *T* y una interrogación *q* , y siendo

*N*= el número total de documentos en la colección

*R*= el número de documentos relevantes para *q*

*n*= el número de documentos que poseen el término

*r* = el número de documentos relevantes que poseen el término

se pueden computar los siguientes tipos de documentos:

| | | |
|---|---|---|
| Documentos relevantes que poseen el término ***r*** | Documentos No relevantes que poseen el término ***n-r*** | Documentos con el término ***n*** |
| Documentos relevantes que no poseen el término ***R-r*** | Documentos No relevantes que no poseen el término ***N-R-n*** | Documentos sin el término ***N-n*** |
| Documentos relevantes ***R*** | Documentos No relevantes ***N-R*** | Total documentos en colección ***N*** |

de allí se pueden derivar cuatro funciones de pesado:

1)

$$p_{iq} = \log \frac{\left(\frac{r}{R}\right)}{\left(\frac{n}{N}\right)}$$

Esta función representa la relación que existe entre la proporción de los documentos relevantes donde ocurre el término *T* y la proporción de toda la colección donde ocurre el mismo término.

2)

$$P_{iq} = \log \frac{\left(\frac{r}{R}\right)}{\left(\frac{n-r}{N-R}\right)}$$

Esta función representa la relación de la proporción de los documentos relevantes frente a los no-relevantes.

3)

$$P_{iq} = \log \frac{\left(\frac{r}{R-r}\right)}{\left(\frac{n}{N-n}\right)}$$

Esta función representa la relación entre la probabilidad de relevancia del término (relación entre el número de documentos relevantes donde ocurre *T,* y el número de documentos relevantes donde no ocurre *T)*; y la probabilidad del término en la colección (relación entre el número total de documentos donde ocurre *T* y el número total de documentos donde no ocurre *T*).

4)

$$P_{iq} = \log \frac{\left(\frac{r}{R-r}\right)}{\left(\frac{n-r}{N-n-R+r}\right)}$$

Esta función representa la relación entre la probabilidad de relevancia de los términos y la probabilidad de su no-relevancia.

# Capítulo 7

## Indización basada en la Semántica Latente

La técnica de Indización basada en la Semántica Latente, conocida en inglés como “Latent Semantic Indexing (LSI)”, es una técnica matemático/estadística que se presenta como una variante del método de recuperación vectorial. Fue propuesta por Dumais, Furnas y Landauer [60] a fin de la década del 80 con la intención de superar el modelo de recuperación basado en la equiparación léxica de las palabras. Propone la organización automática de los textos siguiendo una estructura semántica que estaría dada por el patrón de uso de las palabras a través de los documentos, intentando inferir las relaciones que están parcialmente ocultas en la variabilidad del vocabulario usado en el discurso[1]. El Prof. Juan-Miguel Gracía de la Universidad del País Vasco [61] brinda un ejemplo sencillo del alcance de esta técnica. El sostiene que si se consideran los términos *coche, automóvil, conductor, elefante*, se observa que *coche* y *automóvil* son sinónimos, que *conductor* está relacionado con ambos, pero *elefante* no está relacionado con ninguno. En una búsqueda literal, si se introduce el término *automóvil*, se recuperará solo los documentos que contienen ese término, no se recuperará los que contienen *coche*, ni tampoco los que hablen de *conductores* que también puede ser deseable recuperar, aunque con menor interés. Hay documentos semanticamente similares que no comparten los términos *coche* y *automóvil*, y que pueden tener en común algunos otros términos, por ejemplo *chofer, motor, vehículo, chasis, combustible, parabrisas, neumáticos, Renault, volante, etc.* ; téminos que seguramente no estarán presentes en los documentos que contienen el término *elefante*. Si se supone que en una colección, la palabra *coche* y la palabra *automóvil* ocurren en 100 documentos y que en 95 de esos documentos ocurren juntas, es razonable proponer que la ausencia de la palabra *automóvil* en un documento que contiene *coche* sea considerado como una particularidad que no deba contemplarse y consecuentemente se desee recuperar el documento en respuesta a una interrogación que contenga la palabra *coche.* Con el Análisis Semántico Latente se pretende modelar matemáticamente estas relaciones de asociatividad. Si bien su origen se encuentra en la RI bajo la denominación de Indización basada en la Semántica Latente, fueron los mismos autores quienes posteriormente generalizaron la

[1] En este contexto, “semántica” implica que los términos de un documento pueden ser tomados como referentes del tema que trata el documento.

denominación a Análisis Semántico Latente cuando ampliaron su aplicación a diversos aspectos lingüísticos y cognitivos [62], [63].

Con esta técnica se construye un “espacio semántico” donde los términos y los documentos que están fuertemente asociados son ubicados unos cerca de otros, reflejando los patrones de asociación entre los datos más importantes e ignorando los menos importantes, es decir los que tienen menor influencia y que pueden constituir “ruido” en el momento de la recuperación. Se debe tener en cuenta que el tipo de estructura en el cual cada inferencia de asociatividad puede darse no está limitado a una correlación entre pares, como en este ejemplo sencillo que se acaba de exponer.

La técnica estadística particular que se aplica es la de **descomposición de valores singulares (SVD)** de una matriz. Esta es una técnica ampliamente usada para descomponer una matriz en varias matrices que exhiben las propiedades más importantes de la matriz original. La técnica en sí tiene varios usos, uno de ellos, el que interesa aquí, es la habilidad de partir el espacio vectorial en subespacios de menores dimensiones.

Para explicar en detalle esta técnica se usará el ejemplo que los autores Deerwester, Dumais y Harshman utilizan en un trabajo de 1990 [64]. Parten de una colección de 9 documentos de tipo memorandos, donde 5 corresponden a la temática “interacción hombre-computadora” (c1 a c5) y 4 tratan sobre “gráficos” (m1 a m4). Se crea una matriz X de asociación término-documento, donde el valor numérico representa la cantidad de veces que el término aparece en el documento.

**Titles:**

c1: *Human* machine *interface* for Lab ABC *computer* applications
c2: A *survey* of *user* opinion of *computer system response time*
c3: The *EPS user interface* management *system*
c4: *System* and *human system* engineering testing of *EPS*
c5: Relation of *user*-perceived *response time* to error measurement

m1: The generation of random, binary, unordered *trees*
m2: The intersection *graph* of paths in *trees*
m3: *Graph minors* IV: Widths of *trees* and well-quasi-ordering
m4: *Graph minors*: A *survey*

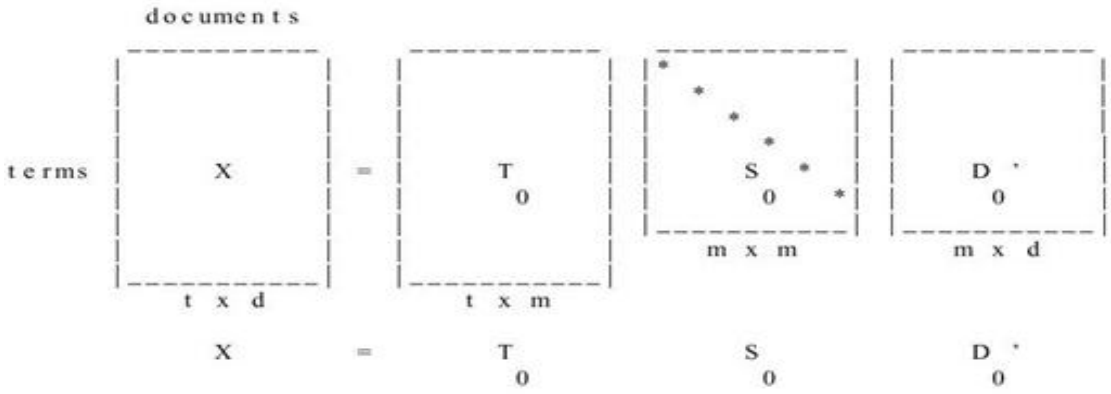
| **Terms** | **Documents** | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | c1 | c2 | c3 | c4 | c5 | m1 | m2 | m3 | m4 |
| *human* | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| *interface* | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| *computer* | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| *user* | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| *system* | 0 | 1 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| *response* | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| *time* | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| *EPS* | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| *survey* | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| *trees* | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| *graph* | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 |
| *minors* | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

Se crea una matriz ***X*** de asociación término-documento, donde el valor numérico en cada celda representa la cantidad de veces que el término aparece en el documento:

$$X = \begin{matrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \end{matrix}$$

Cualquier matriz rectangular[2], por ejemplo *t x d* matriz término documento, ***X***, puede ser descompuesta en el producto de 3 matrices[3]:

$$X_{txd} = To_{txm} * So_{mxm} * Do'_{mxd}$$




[2] Una matriz rectangular es aquella en el que el número de filas es diferente al número de las columnas

[3] El producto de dos matrices A y B está definido cuando el número de columnas de A coincide con el número de filas B. Si A es una matriz *m x p,* B debe ser una matriz *p x n* y la matriz resultado del producto será, entonces, una matriz de *m x n*. Así, sea la matriz A de 3 x 2 y la matriz B de 2 x 3, al multiplicarlas se obtiene una matriz de 2 x 2.

Por ejemplo,

$$A = \begin{pmatrix} 1 & 1 \\ 2 & 2 \\ 2 & 2 \end{pmatrix}, \quad \text{y su transpuesta } A^t = \begin{matrix} 1 & 2 & 2 \\ 1 & 2 & 2 \end{matrix}, \text{ entonces } A^tA = \begin{pmatrix} 9 & 9 \\ 9 & 9 \end{pmatrix}$$

Donde cada elemento de la nueva matriz es el resultado de la suma de los productos de cada uno de los elementos de las columnas de A por los elementos de las filas de $A^t$: (1 . 1) + (2 . 2) + (2 . 2) = 9

Donde ***T0*** es una matriz ortonormal[4] en la que el número de filas corresponde a las filas de X pero sus columnas *m* corresponden a nuevas variables obtenidas de tal manera que no hay ninguna correlación entre dos columnas cualesquiera, esto es, cada una es linealmente independiente de las demás. Sus valores son los vectores singulares o autovectores de $\mathbf{X}\,\mathbf{X}^t$ (que corresponden a los vectores izquierdos de los 12 términos). De igual modo, ***D0*** es también una matriz ortonormal, donde el número de sus columnas corresponden a la matriz original, pero sus filas *m* están compuestas por los vectores singulares o autovectores de $\mathbf{X}^t\,\mathbf{X}$ (que corresponden a los vectores derechos de los 9 documentos). La tercer matriz ***S0***, es una matriz diagonal de *m x m*, cuyos valores son los llamados valores singulares o autovalores y que se encuentran a lo largo de la diagonal central. Cuando las tres matrices son multiplicadas, la matriz original es reconstruida.

Entonces, para el ejemplo, las respectivas matrices son:

$T_0 =$

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.22 | -0.11 | 0.29 | -0.41 | -0.11 | -0.34 | 0.52 | -0.06 | -0.41 |
| 0.20 | -0.07 | 0.14 | -0.55 | 0.28 | 0.50 | -0.07 | -0.01 | -0.11 |
| 0.24 | 0.04 | -0.16 | -0.59 | -0.11 | -0.25 | -0.30 | 0.06 | 0.49 |
| 0.40 | 0.06 | -0.34 | 0.10 | 0.33 | 0.38 | 0.00 | 0.00 | 0.01 |
| 0.64 | -0.17 | 0.36 | 0.33 | -0.16 | -0.21 | -0.17 | 0.03 | 0.27 |
| 0.27 | 0.11 | -0.43 | 0.07 | 0.08 | -0.17 | 0.28 | -0.02 | -0.05 |
| 0.27 | 0.11 | -0.43 | 0.07 | 0.08 | -0.17 | 0.28 | -0.02 | -0.05 |
| 0.30 | -0.14 | 0.33 | 0.19 | 0.11 | 0.27 | 0.03 | -0.02 | -0.17 |
| 0.21 | 0.27 | -0.18 | -0.03 | -0.54 | 0.08 | -0.47 | -0.04 | -0.58 |
| 0.01 | 0.49 | 0.23 | 0.03 | 0.59 | -0.39 | -0.29 | 0.25 | -0.23 |
| 0.04 | 0.62 | 0.22 | 0.00 | -0.07 | 0.11 | 0.16 | -0.68 | 0.23 |
| 0.03 | 0.45 | 0.14 | -0.01 | -0.30 | 0.28 | 0.34 | 0.68 | 0.18 |

$S_0 =$

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 3.34 | | | | | | | | |
| | 2.54 | | | | | | | |
| | | 2.35 | | | | | | |
| | | | 1.64 | | | | | |
| | | | | 1.50 | | | | |
| | | | | | 1.31 | | | |
| | | | | | | 0.85 | | |
| | | | | | | | 0.56 | |
| | | | | | | | | 0.36 |

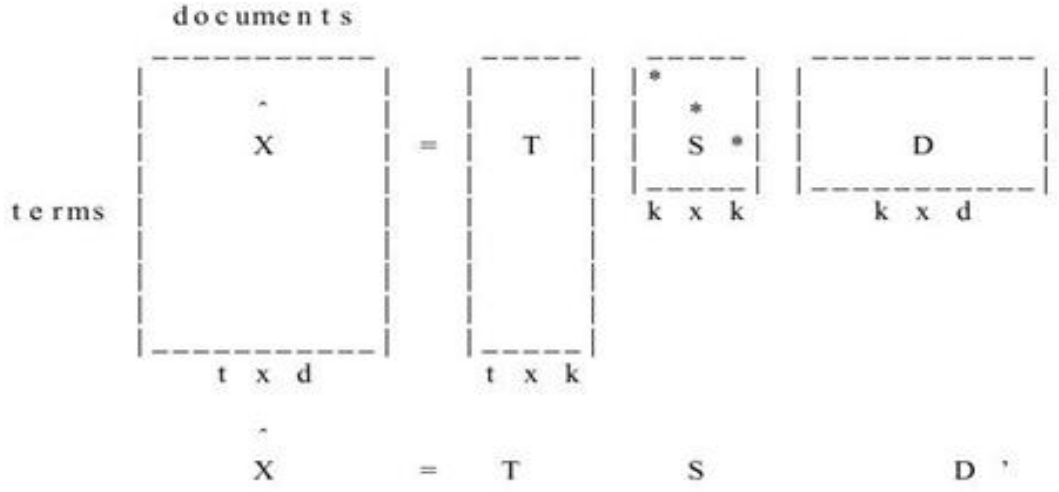
[4] Una matriz ortonormal es aquella que multiplicada por su transpuesta es igual a la matriz identidad. La matriz identidad es una matriz cuadrada cuyos únicos elementos no nulos son los de la diagonal y valen 1.

$D_0$ =

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 0.20 | -0.06 | 0.11 | -0.95 | 0.05 | -0.08 | 0.18 | -0.01 | -0.06 |
| 0.61 | 0.17 | -0.50 | -0.03 | -0.21 | -0.26 | -0.43 | 0.05 | 0.24 |
| 0.46 | -0.13 | 0.21 | 0.04 | 0.38 | 0.72 | -0.24 | 0.01 | 0.02 |
| 0.54 | -0.23 | 0.57 | 0.27 | -0.21 | -0.37 | 0.26 | -0.02 | -0.08 |
| 0.28 | 0.11 | -0.51 | 0.15 | 0.33 | 0.03 | 0.67 | -0.06 | -0.26 |
| 0.00 | 0.19 | 0.10 | 0.02 | 0.39 | -0.30 | -0.34 | 0.45 | -0.62 |
| 0.01 | 0.44 | 0.19 | 0.02 | 0.35 | -0.21 | -0.15 | -0.76 | 0.02 |
| 0.02 | 0.62 | 0.25 | 0.01 | 0.15 | 0.00 | 0.25 | 0.45 | 0.52 |
| 0.08 | 0.53 | 0.08 | -0.03 | -0.60 | 0.36 | 0.04 | -0.07 | -0.45 |

Después de esta tarea de descomposición, conocida como SVD, las *k* dimensiones más importantes, aquellas con los valores singulares más altos en ***S0*** son seleccionadas. Para ello se ordenan los valores singulares de ***S0*** de mayor a menor, se escogen los primeros *k* valores y todos los otros factores son omitidos. Se tiene entonces una nueva matriz singular ***S*** de rango *k.* Al borrar las columnas correspondientes (a las eliminadas de S0) en ***T0*** *y* ***D0*** se obtienen dos matrices reducidas: ***T*** y ***D***. A partir de allí, se está en condiciones de volver a construir la matriz original ***X*** reducida a *k* dimensiones. El nivel de reducción de dimensionalidad que se elige es crítico. Se supone que *k* debería ser bastante grande para mostrar la verdadera estructura en los datos, pero suficientemente pequeño como para no modelar el ruido. En casos reales, generalmente el conjunto *k* se define entre 100 y 300 factores. La solución de la reducción de la dimensionalidad genera entonces un vector de *k* verdaderos valores para representar cada documento. La matriz reducida , que llamaremos ***Xhihat,*** modela de manera confiable a la base de datos contenida en ***X***.

Para el ejemplo de los 9 documentos, la reducción se realiza a *k* = 2



$X\approx$ $T$ $S$ $D'$

```
0.22 -0.11    3.34            0.20 0.61  0.46  0.54 0.28 0.00 0.02 0.02 0.08
0.20 -0.07          2.54     -0.06 0.17 -0.13 -0.23 0.11 0.19 0.44 0.62 0.53
0.24  0.04
0.40  0.06
0.64 -0.17
0.27  0.11
0.27  0.11
0.30 -0.14
0.21  0.27
0.01  0.49
0.04  0.62
0.03  0.45
```

*Xhihat* =

```
 0.16   0.40   0.38   0.47   0.18 -0.05 -0.12 -0.16 -0.09
 0.14   0.37   0.33   0.40   0.16 -0.03 -0.07 -0.10 -0.04
 0.15   0.51   0.36   0.41   0.24  0.02  0.06  0.09  0.12
 0.26   0.84   0.61   0.70   0.39  0.03  0.08  0.12  0.19
 0.45   1.23   1.05   1.27   0.56 -0.07 -0.15 -0.21 -0.05
 0.16   0.58   0.38   0.42   0.28  0.06  0.13  0.19  0.22
 0.16   0.58   0.38   0.42   0.28  0.06  0.13  0.19  0.22
 0.22   0.55   0.51   0.63   0.24 -0.07 -0.14 -0.20 -0.11
 0.10   0.53   0.23   0.21   0.27  0.14  0.31  0.44  0.42
-0.06   0.23 -0.14 -0.27   0.14  0.24  0.55  0.77  0.66
-0.06   0.34 -0.15 -0.30   0.20  0.31  0.69  0.98  0.85
-0.04   0.25 -0.10 -0.21   0.15  0.22  0.50  0.71  0.62
```

Volviendo al ejemplo dado por Gracia, siendo $d_j$ un documento que contiene el término *automóvil* 2 veces y $d_k$ otro documento que contiene el término *coche* 5 veces, una manera de relacionarlos es mostrando los términos con **componentes no nulas que comparten** (*motor, vehícul.o, chasis, conductor, rueda)* mediante el producto escalar de los vectores $d_j$ y $d_k$. Como se sabe, del producto escalar se obtiene como resultado un número, no otro vector, por lo que puede ser utilizado como una medida de distancia semántica entre ambos documentos. El Producto escalar de dos vectores se define como el producto de sus dos módulos por el coseno del ángulo que forman:

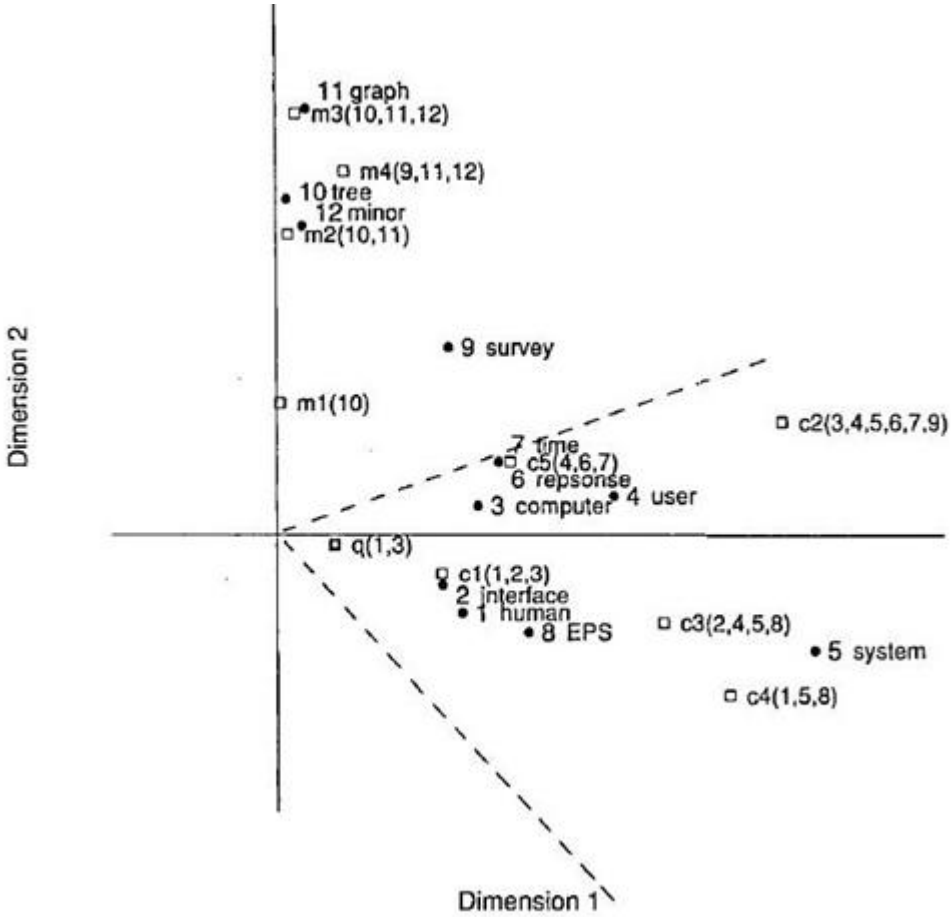
$$\vec{A} * \vec{B} = \left|\vec{A}\right| * \left|\vec{B}\right| \cos \vartheta$$

Cuando los vectores son ortogonales, es decir que forman un ángulo de 90 grados, el coseno vale 0 por lo tanto su producto escalar vale 0. Éste es el caso en el que dos vectores documentales no poseen ningún término en común. Para dos vectores parecidos, el coseno se aproxima a 1. Así el coseno se comporta de manera similar a un índice de correlación. El problema es que de esta manera la medición de la distancia entre dos vectores solo está ponderando la existencia del término, no la

cantidad de veces que aparece. Por ello, entonces, es más conveniente normalizar el valor del producto escalar dividiéndolo por el producto de las normas de los vectores.

En el modelo de Semántica Latente presentado por Deerwester el producto escalar se utiliza para medir la distancia semántica entre los vectores de la matriz ***Xhihat***. El producto escalar entre dos filas brinda la similitud entre dos términos a través de todos los documentos.

Por lo reducido del ejemplo y la cantidad de dimensiones elegidas (2), es factible mostrar su interpretación geométrica. Las filas de las matrices reducidas se toman como cooordenadas de los puntos que representan a los documentos (cuadrados blancos) y a los términos (puntos negros). Los ejes están en escala apropiada en relación con los valores de la diagonal ***S***. Los productos escalares, (coseno del ángulo) establecen la relación de similitud entre los distintos vectores. Para la consulta *q* "human computer interaction", vemos que solo comparte dos términos con la base de datos: "human" (término número 1) y "computer" (término número 3):




En lugar de realizar la representación de los documentos y de las interrogaciones directamente con un conjunto de palabras independientes, con esta técnica se los

representa como valores continuos en cada una de las dimensiones ortogonales de $k$ . Dado que el número de los factores o dimensiones es mucho menor que el número de términos únicos, las palabras no serán independientes. Por ejemplo, si dos términos son usados en contextos similares (documentos) , ellos tendrán vectores similares en la dimensión reducida de la representación LSI. Esta técnica permite capturar mejor la estructura que una correlación término-término o documento-documento.

# PARTE III

# TRABAJO EXPERIMENTAL

# Capítulo 8

## Estudio 1: Ley de Zipf y transición de Goffman

### Objetivo

En este estudio se aplica sobre 3 artículos científicos en idioma inglés el procesamiento textual necesario para:

1) Mostrar lo que sostiene la Ley de rango-frecuencia de Zipf
2) Mostrar el resultado obtenido al aplicar lo que conocemos como Transición de Goffman
3) Mostrar los resultados obtenidos al aplicar diferentes variantes en el tratamiento tetxual del título/resumen por un lado y del texto completo del artículo por el otro. Las variantes principales son:
   a. Tomar la totalidad de las palabras
   b. Aplicar a la totalidad de las palabras un proceso de stemming
   c. Aplicar a los sustantivos un proceso de stemming

### Fuente de datos

Se tomaron 3 artículos publicados en revistas científicas internacionales de reconocido renombre en el área de la óptica. Los 3 trabajos corresponden a la temática de "*cristales fotorefractivos*" y fueron publicados en las revistas Optik, Optics Communication y Optical Engineering en el año 2000[1]. Los 3 artículos están escritos en idioma inglés y su extensión y características generales se describen en la TABLA I

| | **Opt. Comm.** | **Opt. Eng.** | **Optik** |
|---|---|---|---|
| Pags. (Word doble esp) | 18 | 20 | 21 |
| Cant. Palabras | 1261 | 3190 | 3676 |
| Cant.Palabras dif. | 451 | 578 | 737 |
| Cant. Figuras | 6 | 5 | 8 |
| Cant. Fórmulas | 4 | 5 | 28 |
| Cant.Referencias | 14 | 14 | 13 |

TABLA I

[1] Disponibles en el CD en \Tesina\Experimental\Estudio1\ nombrados como t01.doc,t02.doc y t03.doc .

## Procesamiento de los textos

Se realizó sobre cada uno de los textos digitales de los artículos una serie de tareas utilizando diferentes herramientas computacionales con la finalidad de obtener los resultados propuestos en los objetivos. Se resume a continuación la secuencia seguida :

**Paso 1**: Se realizó la limpieza de los textos. Se procedió a borrar

- imágenes
- gráficos
- tablas
- información de pertenencia
- bibliografía
- notas al pie
- aclaraciones en el texto del tipo: fig., graf., llamadas, etc.

**Paso 2:** Se separó el título y el resumen por un lado y el texto propiamente dicho por el otro[2].

**Paso 3:** Para cada tipo de texto generado en el paso anterior se calculó las frecuencias de las palabras utilizando el programa Wlist[3].

**Paso 4:** Para cada tipo de texto generado en el Paso 2 se calculó las frecuencias de las palabras resultantes de aplicar un proceso de stemming[4]. Para realizar el stemming se utilizó el programa KTERM. Para el cálculo de frecuencias el programa WLIST.

---

[2] Archivos *t01_A.txt* y *t01_F.txt*, etc. disponibles en el CD en \Tesina\Experimental\Estudio1\Abstracts y \Tesina\Experimental\Estudio1\Full, respectivamente.

[3] Archivos *t01_A_W.txt, t01_F_W.txt,* etc. Disponibles en el CD en \Tesina\Experimental\Estudio1\Abstracts\A_Compl y \Tesina\Experimental\Estudio1\Full\F_Compl. Información sobre el programa WLIST en \Tesina\Experimental\Programas

[4] Archivos *t01_A_K2.txt, t01_F_K2.txt* etc. Disponibles en el CD en \Tesina\Experimental\Estudio1\Abstracts\A_Stemm, \Tesina\Experimental\Estudio1\Full\F_Stemm. Información sobre el programa KTERM en \Tesina\Experimental\Programas

**Paso 5:** Para cada tipo de texto generado en el Paso 2 se extrajeron los sustantivos, se les aplicó un proceso de sttemming y se calculó las frecuencias de aparición[5]. Para realizar el stemming y la selección de los sustantivos se utilizó el programa KTERM. Para el cálculo de frecuencias el programa WLIST.

**Paso 6**: Se importaron los archivos obtenidos en el paso 3, a EXCEL. Se ordenaron las palabras de la más frecuente a la menos frecuente, se le asignó a cada palabra un rango y se calculó la constante de Zipf multiplicando el rango por la frecuencia. Se graficó la curva rango/frecuencia en escala logarítmica. Se aplicó la línea de tendencia y se calculó la ecuación de la recta[6].

**Paso 7**: Se importaron los archivos obtenidos en el paso 3, 4 y 5 a EXCEL y se calculó la Transición de Goffman. Para ello se ordenaron las palabras por la frecuencia y se contaron las palabras con frecuencia 1. Se aplicó la formula: =(-1+SQRT(1+8* [cant. Pal.Frec. 1]))/2[7].

## Resultados

Se resumen aquí los resultados obtenidos. Se realiza un análisis general tratando de marcar los principales rasgos

[5] En CD: \Tesina\Experimental\Estudio1\Abstracts\A_Stemm_Noun\ ***t01_A_K.txt,*** etc.
y \Tesina\Experimental\Estudio1\Full\F_Stemm_Noun\ ***t01_F_K.txt***

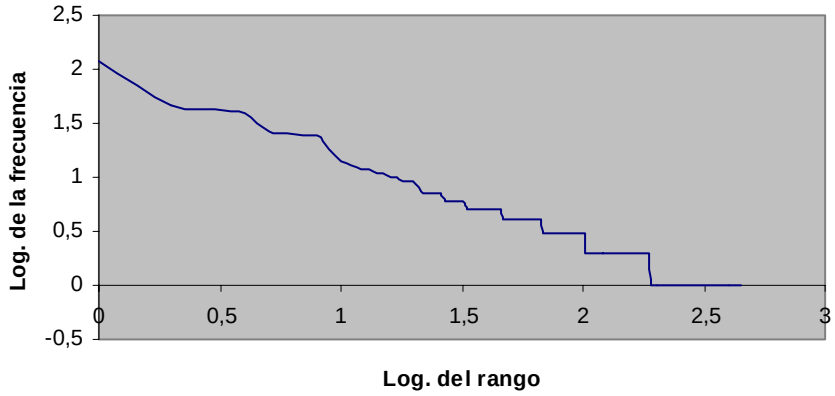
[6] En CD: \Tesina\Experimental\Estudio1\Full\F_Compl\ ***t01_F_W_Zipf.xls,*** etc.

[7] En CD: \Tesina\Experimental\Estudio1\Abstracts\A_Compl \ ***t01_A_W_Goff.xls,*** etc.
\Tesina\Experimental\Estudio1\Abstracts\A_Stemm\ ***t01_A_K2_Goff.xls,*** etc.
\Tesina\Experimental\Estudio1\Abstracts\A_Stemm_Noun\ ***t01_A_K_Goff.xls***, etc.
\Tesina\Experimental\Estudio1\Full\F_Compl\ ***t01_F_W_Goff.xls,*** etc.
\Tesina\Experimental\Estudio1\Full\F_Stemm\ ***t01_F_K2_Goff.xls,*** etc.
\Tesina\Experimental\Estudio1\Full\F_Stemm_Noun\. ***t01_F_K_Goff.xls,*** etc.

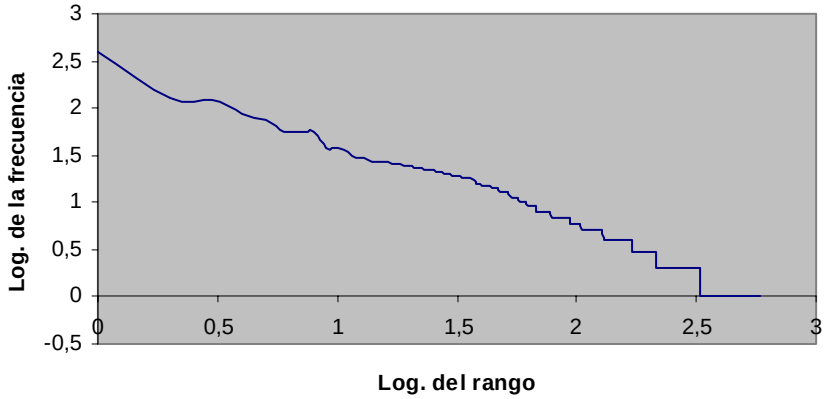
## Ley rango/frecuencia de Zipf

Se observa que la curva resultante es de similares características en los tres textos.

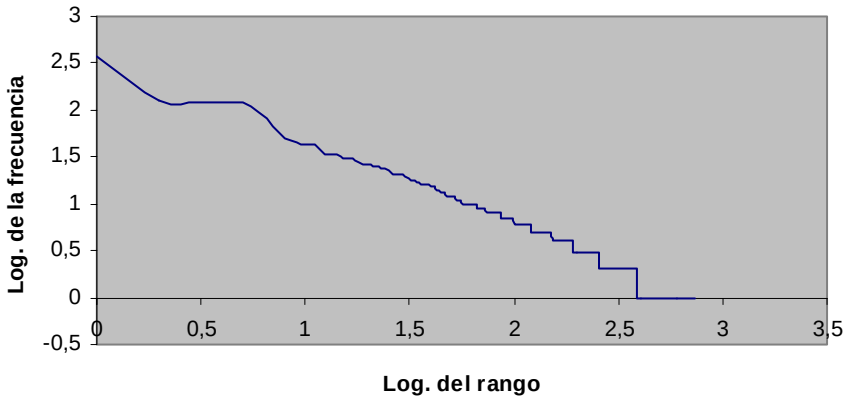
Rango/frecuencia (Texto 01)



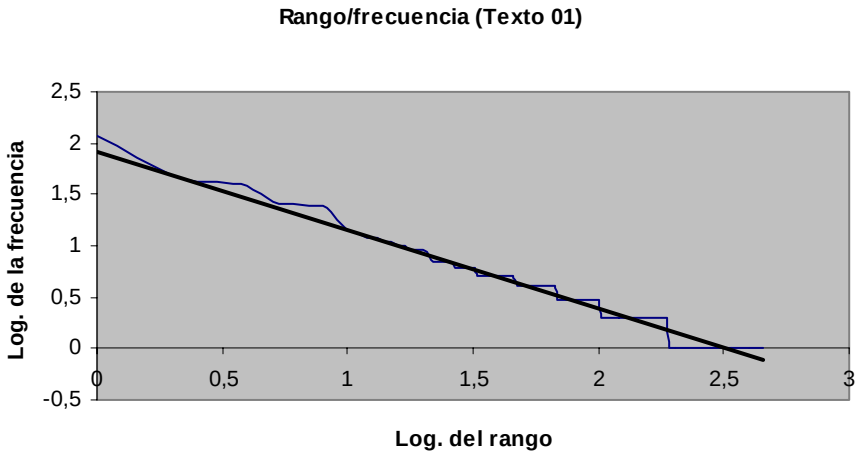

Rango/frecuencia (Texto 02)



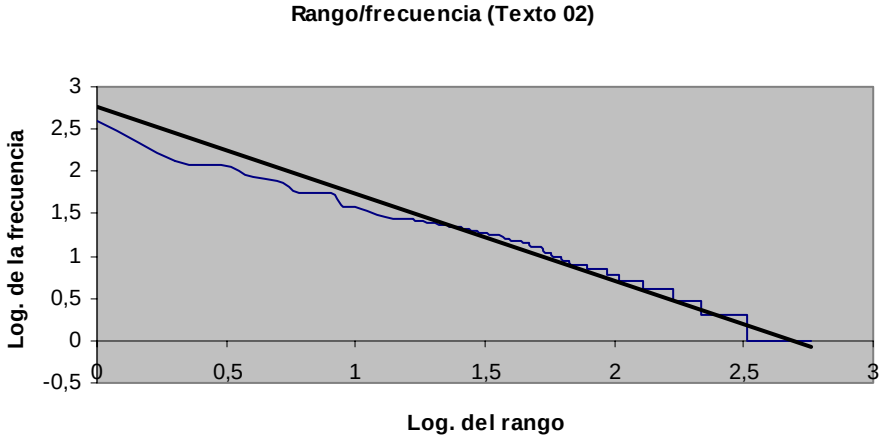

**Rango/frecuencia (Texto 03)**



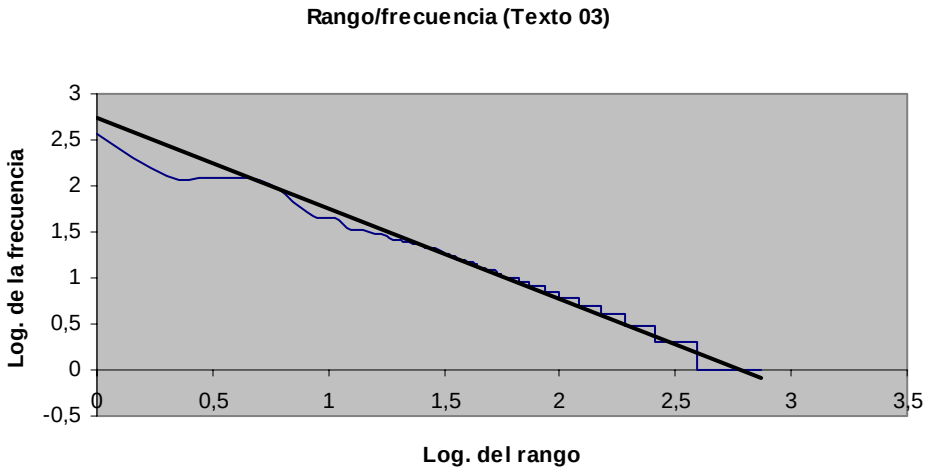
Analizando con mayor detalle vemos que las líneas de tendencia nos muestran curvas de pendiente negativa, cercana a –1, con valores similares en la ordenada al origen y coeficiente de correlación:




| | |
|---|---|
| Pendiente | -0,7647 |
| Ordenada al origen | 1,9227 |
| Coeficiente de correlación | 0,951 |




| | |
|---|---|
| Pendiente | -1,0305 |
| Ordenada al origen | 2,7751 |
| Coeficiente de correlación | 0,9679 |




| | |
|---|---|
| Pendiente | -0,9887 |
| Ordenada al origen | 2,7475 |
| Coeficiente de correlación | 0,9696 |

## Transición de Goffman

W. Goffman desarrolló un método para seleccionar los términos más importantes de un texto que fue documentado por Miranda Lee Pao [65] y se lo conoce como la Transición de Goffman.

El observó que la condición bajo la cual la Ley de Zipf opera es solo si se considera la ocurrencia de palabras de alta frecuencia. Estas ocupan una única posición en el rango de toda la distribución de palabras. Esto es, para las palabras de alta frecuencia en un texto dado, no hay dos palabras con la misma frecuencia. Es decir, una y solo una palabra ocurre con la mayor frecuencia y tiene el rango 1, una y solo una palabra es segunda en frecuencia y tiene el rango 2 y así sucesivamente.

En cambio, si se revisan las palabras de baja frecuencia se observa que varias palabras distintas tienen el mismo valor de frecuencia. Estos comportamientos diferentes entre las palabras de alta frecuencia y las de baja, describen y predicen la distribución de cualquier texto.

Goffman sostiene, entonces, que es razonable esperar una región crítica en la cual la transición del comportamiento de las palabras desde las frecuencias altas a las bajas tiene lugar y que serán dichas palabras las de mayor significación en cualquier texto.

Los resultados de aplicar la formula de Goffman (paso 7) a los 3 textos seleccionados en sus diferentes variantes se detallan en las tablas siguientes:

**Texto 1 – Optics Communications**

**Título**

Phase-stepping technique with an electro-optic crystal in digital speckle pattern interferometry

**Resumen**

The use of an electro-optic crystal as phase-stepping device in a digital speckle interferometer is discussed. Phase stepping is introduced by varying the external voltage applied to the crystal. The crystal calibration procedure is outlined and errors are discussed. Experimental results compared to those obtained with a piezo-electrically driven mirror are presented.

**Palabras claves de los autores:** BSO crystal, phase stepping, metrology, digital speckle pattern interferometry

**Título/Resumen-Completo:**

| | |
|---|---|
| **Crystal** | **4** |
| **Stepping** | **3** |
| **Phase** | **3** |

Goffman= 7,13

**Título/Resumen-Stemming:**

| | |
|---|---|
| **Crystal** | **4** |
| **Phase** | **3** |
| **Step** | **3** |

Goffman= 6,00

**Título/Resumen-Sustantivo+Stem:**

| | |
|---|---|
| **Crystal** | **4** |
| **Phase** | **3** |
| **Step** | **3** |

Goffman= 4,62

**Total-Completo:**

| | |
|---|---|
| **Phase** | **25** |
| **Crystal** | **18** |

Goffman= 22,27

**Total-Stemming:**

| | |
|---|---|
| **Phase** | **25** |
| **Crystal** | **19** |
| **Fringe** | **15** |
| **Step** | **12** |
| **Beam** | **10** |

*[1)] **stepping (5) steps (4) step (2) stepper (1)**

Goffman= 17,11

**Total- Sustantivo + Stem:**

| | |
|---|---|
| **Phase** | **25** |
| **Crystal** | **19** |
| **Fringe** | **15** |
| **Step** | **12** |
| **Beam** | **10** |
| **Voltage** | **10** |
| **Pattern** | **7** |
| **Speckle** | **6** |

Goffman= 13,58

**Texto 2 – Optical Engineering**

**Título**

Fringe visibility analysis with different scale apertures in speckle photography

**Resumen**

The use of different scale aperture pupils for image recording in speckle photography is analysed. In particular a double-exposure specklegram is considered. The ensemble-average intensity in the Fourier plane is analytically derived and fringe visibility is investigated. The theoretical results are verified by in-plane displacement translation experiments.

**Palabras claves del autor:** Speckle photography, scales aperture

**Título/Resumen-Completo:**

| | |
|---|---|
| **Plane** | **2** |
| **Visibility** | **2** |
| **Speckle** | **2** |
| **Scale** | **2** |
| **Photography** | **2** |
| **Fringe** | **2** |

*[1)] **Apertures (1) Aperture (1)**

Goffman= 7,64

**Título/Resumen-Stemming:**

| | |
|---|---|
| **Aperture** | **2** |
| **Plane** | **2** |
| **Visibility** | **2** |
| **Speckle** | **2** |
| **Scale** | **2** |
| **Photograph** | **2** |
| **Fringe** | **2** |

Goffman= 6,15

**Título/Resumen-Sustantivo+Stem:**

| | |
|---|---|
| **Aperture** | **2** |
| **Plane** | **2** |
| **Photograph** | **2** |
| **Visibility** | **2** |
| **Speckle** | **2** |
| **Scale** | **2** |
| **Fringe** | **2** |

Goffman= 4,62

**Total-Completo:**

| | |
|---|---|
| **Plane** | **34** |
| **Exposures** | **30** |
| **Speckle** | **29** |
| **Pupil** | **27** |
| **Difraction** | **27** |
| **Intensity** | **26** |
| **Aperture** | **25** |
| **Visibility** | **23** |
| **Displacement** | **22** |
| **Scale** | **21** |
| **Pupils** | **19** |
| **Apertures** | **19** |
| **Fringe** | **19** |
| **Fringes** | **18** |
| **Fourier** | **18** |
| **Image** | **18** |
| **Specklegram** | **14** |
| **Exposures** | **13** |
| **Images** | **12** |
| **Speckles** | **10** |

*[1)] **Scales (1) Displacement (1) Intensities (2)**

Goffman= 21,73

**Total-Stemming:**

| | |
|---|---|
| **Pupil** | **46** |
| **Aperture** | **44** |
| **Exposure** | **43** |
| **Speckle** | **40** |
| **Fringe** | **37** |
| **Plane** | **34** |
| **Image** | **30** |
| **Diffract** | **28** |
| **Intense** | **28** |
| **Scale** | **27** |
| **Visibily** | **23** |
| **Pattern** | **16** |
| **Halo** | **13** |
| **Correlate** | **10** |

Goffman= 14,87

**Total- Sustantivo + Stem:**

| | |
|---|---|
| **Pupil** | **46** |
| **Aperture** | **44** |
| **Exposure** | **43** |
| **Speckle** | **39** |
| **Fringe** | **37** |
| **Plane** | **34** |
| **Image** | **30** |
| **Intense** | **28** |
| **Diffract** | **27** |
| **Visibility** | **23** |
| **Scale** | **22** |
| **Pattern** | **16** |
| **Halo** | **13** |
| **Correlate** | **9** |
| **Diffuse** | **8** |
| **Spot** | **8** |
| **Lens** | **6** |
| **Axis** | **5** |
| **Figure** | **5** |
| **Photograph** | **5** |
| **Surface** | **5** |
| **Interfere** | **4** |
| **Light** | **4** |

Goffman= 11,00

**Texto 3 – Optik**

**Título**

Talbot and Lau phenomena implementation via a photorefractive converter

**Resumen**

We propose a novel implementation of the Lau and Talbot effects. It employs a grating registered as modulation of birefringence in a Bi12SiO20 (BSO) crystal. The systems proposed allow to control the visibility and in particular to reverse the contrast of the self images and Lau fringes by rotating an analyzer. A theoretical approach is outlined and the predicted behavior is confirmed by results obtained under different experimental conditions. We have taken advantage of the arrangements to demonstrate some applications.

**Palabras claves del autor:** Self-imaging, photorefractive materials, birefringence

**Título/Resumen-Completo:**

| | |
|---|---|
| **Talbot** | **3** |
| **Lau** | **2** |

Goffman= 9,51

**Título/Resumen-Stemming: Ninguna**

Goffman= 8,35

**Título/Resumen-Sustantivo+Stem: Ninguna**

Goffman= 6,30

**Total-Completo:**

| | |
|---|---|
| **Crystal** | **51** |
| **Grating** | **44** |
| **Plane** | **35** |
| **Light** | **28** |
| **Lau** | **25** |
| **Intensity** | **24** |
| **Self** | **21** |
| **Talbot** | **18** |
| **Amplitude** | **18** |
| **Pattern** | **15** |
| **Polarizer** | **10** |
| **Polarized** | **10** |
| **Polarization** | **10** |

*[1)] **Patterns (2) Intensities (2Planes (6) Gratings (10) Crystals (1)**

Goffman= 25,69

**Total-Stemming:**

| | |
|---|---|
| **Crystal** | **52** |
| **Plane** | **41** |
| **Polar** | **37** |
| **Field** | **30** |
| **Light** | **28** |
| **Intense** | **26** |
| **Image** | **24** |
| **Distribute** | **22** |
| **Self** | **21** |
| **Period** | **21** |
| **Amplitude** | **18** |
| **Fringe** | **18** |
| **Contrast** | **17** |
| **Pattern** | **17** |
| **Lens** | **11** |
| **Diffract** | **10** |
| **Axe** | **9** |
| **Phase** | **8** |
| **Dark** | **8** |
| **Optic** | **8** |
| **Interferometer** | **7** |
| **Visibility** | **7** |
| **Focal** | **7** |

Goffman= 17,84

**Total- Sustantivo + Stem:**

| | |
|---|---|
| **Crystal** | **52** |
| **Plane** | **41** |
| **Polar** | **37** |
| **Field** | **30** |

Goffman= 13,37

En todos los casos se observa que el valor de Goffman disminuye cuanto mayor es el procesamiento textual y por lo tanto el vocabulario disminuye.

En el caso de los títulos/resumen los valores de frecuencias más altos son bastante menores que el valor obtenido al calcular Goffman, por lo cual se concluye que no es aplicable esta técnica en textos tan breves.

En los textos completos, considerando los términos con frecuencia ½ valor de Goffman por arriba y por debajo, es una técnica aplicable que pierde efectividad cuanto mayor es el procesamiento al que se somete el texto.

# Capítulo 9

## Estudio 2: Ponderación y modelo del Espacio Vectorial

### Objetivo

En este estudio se aplica sobre 10 registros y 3 interrogaciones de la colección de prueba Cranfield, el procesamiento textual necesario para:

1) Mostrar el funcionamiento básico del modelo de recuperación de información vectorial
2) Mostrar los resultados obtenidos al aplicar dos funciones de similitud diferentes: producto escalar y coseno
3) Mostrar los resultados obtenidos al aplicar diferentes variantes en el tratamiento textual
    a. Aplicar a la totalidad de las palabras un proceso de filtrado con un archivo de “stopword”
    b. Aplicar a la totalidad de las palabras del punto a. un proceso de stemming basado en el algoritmo de Porter
    c. Aplicar a la totalidad de las palabras del punto b. el tipo de ponderación conocido como IDF

### Fuente de datos

Los Proyectos Cranfield I y Cranfield II fueron los primeros experimentos en Recuperación de Información. El Proyecto Cranfield I investigó el comportamiento de 4 sistemas de indización diferentes: la Clasificación Decimal Universal, una clasificación facetada, un catálogo temático alfabético y un sistema post-coordinado de unitérminos. Como resultado arrojó que en términos de “performance” todos los sistemas operaban con los mismos niveles de eficiencia en términos de recuperación. El Proyecto Cranfield II fue diseñado para investigar cada uno de los lenguajes de indización por separado y en todas las combinaciones prácticas posibles y tratar de medir el efecto que cada uno de los aspectos analizados tenía en el rendimiento final del sistema de RI.

Los proyectos Cranfield fueron fundacionales en lo referente a la investigación sobre Recuperación de Información, fundamentalmente por dos cuestiones: 1- el

haber desarrollado las primeras medidas de evaluación (precisión y exhaustividad), 2 – el haber utilizado una colección de documentos e interrogaciones con sus correspondientes juicios de relevancia sobre la cual aplicar de manera sistemática los diferentes test de evaluación.

### El corpus de Cranfield II

La colección utilizada en los test de Cranfield II consiste en 1400 citas con resúmenes de artículos de investigación en el campo temático de la aerodinámica. Está acompañada de un conjunto de 221 preguntas que fueron elaboradas por investigadores de la misma área temática. Además se determinó la relevancia de cada documento de la colección respecto a cada una de las preguntas[1]. La decisión de relevancia fue dada por la misma persona que había originado la pregunta. El juicio de relevancia fue expresado en una escala de 1 a 4 de acuerdo a los siguientes criterios:

a) Asignar puntaje 1 a aquellas referencias que responden completamente la pregunta

b) Asignar puntaje 2 a aquellas referencias con alto grado de relevancia, de manera que sin ellas, la investigación seria impracticable o solo se podría sacar resultados con mucho trabajo extra.

c) Asignar puntaje 3 a aquellas referencias que son muy útiles como "background" general del trabajo o sugieren métodos que permiten abordar ciertos aspectos del trabajo.

d) Asignar puntaje 4 a aquellas referencias con un mínimo de interés, por ejemplo, aquellas que incluyen puntos de vistas históricos.

A los fines del presente trabajo solo se tomó un ejemplo de 10 registros y 3 interrogaciones con los correspondientes juicios de relevancia.

## Procesamiento de los textos

Se realizó sobre los registros seleccionados de la colección una serie de tareas utilizando diferentes herramientas computacionales con la finalidad de obtener los resultados propuestos en los objetivos. Se resume a continuación la secuencia seguida:

[1] Archivos ***cran.all.1400, cran.qry y cranquel*** disponibles en el CD en \Tesina\Experimental\Estudio 2\Cranfield

**Paso 1**: Se obtuvo la matriz término/documento[2] de los 10 primeros registros de la colección Cranfield. Para ello se utilizó una librería desarrollada en MatLab por D. Zeimpekis y E. Gallopoulos de la Universidad de Patras en Grecia llamadas TMG (Ver ANEXO III).

**Paso 2**: Se obtuvo la matriz término/interrogación[3] de 3 preguntas de la colección que estuvieran relacionadas con los 10 documentos seleccionados según los juicios de relevancia de los expertos. Las interrogaciones son las número 3, 65 y 67 de la colección Cranfield utilizando la misma herramienta del paso anterior.

**Paso 3**: Se exportó la matriz de MatLab a Excel. Se realizó el Producto Escalar[4] de cada uno de los vectores de la interrogaciones con cada uno de los vectores documentales.

**Paso 4**: Se aplicó la función de Similitud del Coseno entre cada vector documental y cada vector de interrogación[5]:

$$SimCoseno(d_i, q_j) = \frac{\sum_{k=1}^{n} (x_{ik} * y_{jk})}{\sqrt{\sum_{k=1}^{n} (x_{ik})^2 * \sum_{K=1}^{n} (y_{jk})^2}}$$

**Paso 5**: Se utilizó la herramienta TMG para generar las diferentes variantes: con stemming y ponderación IDF.

---

[2] Archivo ***Estudio 2.xls/Hoja: Matriz_Documento (cran_10)*** disponibles en el CD en \Tesina\Experimental\Estudio 2

[3] Archivo ***Estudio 2.xls/Hoja: Matriz_Query*** disponibles en el CD en \Tesina\Experimental\Estudio 2

[4] Archivo ***Estudio 2.xls/Hoja: Matriz_EscalarQ3, Matriz_EscalarQ65 y Matriz_EscalarQ67*** disponibles en el CD en \Tesina\Experimental\Estudio 2

[5] Archivo ***Estudio 2.xls/Hoja: Matriz_CosenoQ3, Matriz_CosenoQ65 y Matriz_CosenoQ67*** disponibles en el CD en \Tesina\Experimental\Estudio 2

## Resultados

Habiendo aplicado los procesos antes detallados sobre los 10 registros que se detallan a continuación

.I 1
.T
experimental investigation of the aerodynamics of a
wing in a slipstream .
.W
experimental investigation of the aerodynamics of a wing in a slipstream . an experimental study of a wing in a propeller slipstream was made in order to determine the spanwise distribution of the lift increase due to slipstream at different angles of attack of the wing and at different free stream to slipstream velocity ratios . the results were intended in part as an evaluation basis for different theoretical treatments of this problem . the comparative span loading curves, together with supporting evidence, showed that a substantial part of the lift increment produced by the slipstream was due to a /destalling/ or boundary-layer-control effect . the integrated remaining lift increment, after subtracting this destalling lift, was found to agree well with a potential flow theory . an empirical evaluation of the destalling effects was made for the specific configuration of the experiment .

.I 2
.T
simple shear flow past a flat plate in an incompressible fluid of small viscosity .
.W
simple shear flow past a flat plate in an incompressible fluid of small viscosity . in the study of high-speed viscous flow past a two-dimensional body it is usually necessary to consider a curved shock wave emitting from the nose or leading edge of the body . consequently, there exists an inviscid rotational flow region between the shock wave and the boundary layer . such a situation arises, for instance, in the study of the hypersonic viscous flow past a flat plate . the situation is somewhat different from prandtl's classical boundary-layer problem . in prandtl's original problem the inviscid free stream outside the boundary layer is irrotational while in a hypersonic boundary-layer problem the inviscid free stream must be considered as rotational . the possible effects of vorticity have been recently discussed by ferri and libby . in the present paper, the simple shear flow past a flat plate in a fluid of small viscosity is investigated . it can be shown that this problem can again be treated by the boundary-layer approximation, the only novel feature being that the free stream has a constant vorticity . the discussion here is restricted to two-dimensional incompressible steady flow .

.I 3
.T
the boundary layer in simple shear flow past a flat plate .
.W

the boundary layer in simple shear flow past a flat plate .the boundary-layer equations are presented for steady incompressible flow with no pressure gradient .

.I 4
.T
approximate solutions of the incompressible laminar boundary layer equations for a plate in shear flow .
.W
approximate solutions of the incompressible laminar boundary layer equations for a plate in shear flow . the two-dimensional steady boundary-layer problem for a flat plate in a shear flow of incompressible fluid is considered . solutions for the boundary-layer thickness, skin friction, and the velocity distribution in the boundary layer are obtained by the karman-pohlhausen technique . comparison with the boundary layer of a uniform flow has also been made to show the effect of vorticity .

.I 5
.T
one-dimensional transient heat conduction into a double-layer slab subjected to a linear heat input for a small time internal .
.W
one-dimensional transient heat conduction into a double-layer slab subjected to a linear heat input for a small time internal . analytic solutions are presented for the transient heat conduction in composite slabs exposed at one surface to a triangular heat rate . this type of heating rate may occur, for example, during aerodynamic heating .

.I 6
.T
one-dimensional transient heat flow in a multilayer slab .
.W
one-dimensional transient heat flow in a multilayer slab . in a recent contribution to the readers' forum wassermann gave analytic solutions for the temperature in a double layer slab, with a triangular heat rate input at one face, insulated at the other, and with no thermal resistance at the interface . his solutions were for the three particular cases.i propose here to give the general solution to this problem, to indicate briefly how it is obtained using the method of reference 2, and to point out that the solutions given by wassermann are incomplete for times longer than the duration of the heat input .

.I 7
.T
the effect of controlled three-dimensional roughness on boundary layer transition at supersonic speeds .
.W
the effect of controlled three-dimensional roughness on boundary layer transition at supersonic speeds . experiments were performed in the 12-in. supersonic wind tunnel of the jet propulsion laboratory of the california institute of technology to investigate the effect of three-dimensional roughness elements (spheres) on boundary-layer transition on a tained at local mach numbers of 1.90, 2.71, and 3.67 by

varying trip size, position, spacing, and reynolds number per inch . the results indicate that (1) transition from laminar to turbulent flow induced by three-dimensional roughness elements begins when the double row of spiral vortices trailing each element contaminates and breaks down the surrounding field of vorticity, (2) transition appears rather suddenly, becoming more violent with increasing roughness height relative to the boundary-layer thickness, (3) after the breakdown of the vorticity field, the strength of the spiral vortices may still persist in the sublayer of the ensuing turbulent flow, (4) lateral spacing of roughness elements has little effect upon the initial breakdown (contamination) of the laminar flow, and (5) the trip reynolds number where u and v are the velocity and kinematic viscosity at the outer edge of the boundary layer and k is roughness height, such that transition occurs at the roughness position, varies as the position reynolds number to the one-fourth power, viz., where x is trip position .

.I 8
.T
measurements of the effect of two-dimensional and three-dimensional roughness elements on boundary layer transition .
.W
measurements of the effect of two-dimensional and three-dimensional roughness elements on boundary layer transition . in his study of the effect of roughness on transition, h. l. dryden found, on the basis of available data, that the effect of a two-dimensional roughness element such as a /trip wire/ could be represented reasonably well in terms of a functional relation between and, where is the reynolds number of transition based on distance from the leading edge, is the height of the roughness element, and is the boundary-layer displacement thickness at the position of the element . at his suggestion some additional data were obtained, primarily to extend the range to higher values of, during the course of an investigation of transition on a flat plate conducted at the national bureau of standards . after the results on the two-dimensional roughness elements were obtained, it appeared to be desirable to see whether a row of three-dimensional roughness elements would behave in the same way .

.I 9
.T
transition studies and skin friction measurements on an insulated flat plate at a mach number of 5.8 .
.W
transition studies and skin friction measurements on an insulated flat plate at a mach number of 5.8 .
an investigation of transition and skin friction on an insulated flat plate, 5 by 26 in., was made in the galcit 5 by 5 in. hypersonic wind tunnel at a nominal mach number of 5.8 . the phosphorescent lacquer technique was used for transition detection and was found to be in good agreement with total-head rake measurements along the plate surface and pitot boundary-layer surveys . it was found that the boundary layer was laminar at reynolds numbers of at least 5 x 10 . transverse contamination caused by the turbulent boundary layer on the tunnel sidewall originated far downstream of the flat plate leading edge at reynolds numbers of 1.5 to 2 x 10, and spread at a uniform angle of 5 compared to 9 degree in low-speed flow . the effect of two-dimensional and local disturbances was investigated . the technique of air injection into the boundary layer as a means of hastening transition was extensively used . although the onset of transition occurred at reynolds numbers as low as 10, a fully developed turbulent boundary layer was not obtained at reynolds numbers much below 2 x 10 regardless of the amount of air injected . a qualitative discussion of these results is given with emphasis on the possibility of a greater stability of the laminar

boundary layer in hypersonic flow than at lower speeds . direct skin-friction measurements were made by means of the floating element technique, over a range of reynolds numbers verified as being laminar over the complete range . with air injection, turbulent shear was obtained only for reynolds numbers greater than 2 x 10, this value being in good agreement with earlier results of this investigation . the turbulent skin-friction coefficient was found to be approximately 0.40 of that for incompressible flow for a constant value of r, and 0.46 for an effective reynolds number between 5 and 6 x 10 .

.I 10
.T
the theory of the impact tube at low pressure .
.W
the theory of the impact tube at low pressure . a theoretical analysis has been made for an impact tube of the relation between free-stream mach number and the impact and free-stream pressures and densities for extremely low pressures . it is shown that the results differ appreciably from the corresponding continuum relations .

y sobre las 3 interrogaciones siguientes:

**Query 3**

.W
what problems of heat conduction in composite slabs have been solved so far .

Query 65

.W
does the boundary layer on a flat plate in a shear flow induce a pressure gradient .

Query 67

W.
can series expansions be found for the boundary layer on a flat plate in a shear flow .

se han obtenido los siguientes resultados:

**Para el Query 3**

| ESCALAR | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 11 | 4 | 0 | 0 | 0 | 0 |
| | | | | Doc 5 | Doc 6 | | | | |
| COSENO | | | | | | | | | |
| 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,5220 | 0,2097 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| | | | | Doc 5 | Doc 6 | | | | |

| Relevancia | Cranfield | |
|---|---|---|
| query | documento | juicio relev. |
| 3 | 5 | 3 |
| 3 | 6 | 3 |

Donde se observa que los documentos recuperados coinciden con los seleccionados por los expertos.

**Para el Query 65**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ESCALAR | | | | | | | | | |
| 3 | 28 | 17 | 23 | 2 | 3 | 13 | 8 | 25 | 2 |
| | Doc 2 | Doc 3 | Doc 4 | | | | | Doc 9 | |
| | | | | | | | | | |
| COSENO | | | | | | | | | |
| 0,0846 | 0,5385 | 0,7633 | 0,6469 | 0,0671 | 0,1112 | 0,2453 | 0,1837 | 0,3937 | 0,0845 |
| | Doc 2 | Doc 3 | Doc 4 | | | | | | |

| Relevancia | Cranfield | |
|---|---|---|
| query | documento | juicio relev. |
| 65 | 2 | 2 |
| 65 | 3 | 2 |
| 65 | 4 | 3 |

Aquí se observa que los 3 documento seleccionados por los expertos resultan recuperados por el sistema utilizando cualquiera de las dos funciones de similitud. Puede observarse, sin embargo, que el orden que se le asignarían a los documentos según los valores arrojados es diferente: la función del Producto Escalar ordenaria los registros 2 , 4, 3; mientras que la función del coseno lo hace 3, 4, 2. También se observa una fuerte diferencia con respecto al Documento 9 que no fue seleccionado por los expertos para responder a esta interrogación y que, sin embargo el Producto Escalar los recupera con un valor muy alto. Se observa que para este mismo documento, la función del coseno se comportaría mejor.

**Para el Query 67**

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ESCALAR | | | | | | | | | |
| 4 | 28 | 15 | 23 | 2 | 3 | 13 | 9 | 28 | 0 |
| | Doc 2 | Doc 3 | Doc 4 | | | | | Doc 9 | |
| | | | | | | | | | |
| COSENO | | | | | | | | | |
| 0,1207 | 0,5756 | 0,7200 | 0,6916 | 0,0717 | 0,1189 | 0,2623 | 0,2210 | 0,4714 | 0,0000 |
| | Doc 2 | Doc 3 | Doc 4 | | | | | Doc 9 | |

| Relevancia | Cranfield | |
|---|---|---|
| query | documento | juicio relev. |
| 67 | 2 | 1 |
| 67 | 3 | 1 |
| 67 | 4 | 3 |

También puede observarse aquí algo similar a lo que sucedió con el caso anterior. Los 3 documentos seleccionados por los expertos resultarían recuperados por el sistema usando cualquiera de las dos funciones de similitus. Sin embargo, sucede nuevamente algo parecido con el Documento 9 y la diferencia en la escala del juicio de relevancia que coloca al documento 4 claramente en segundo lugar, no se reflejaría en el orden que producirían las funciones.

Aplicando las diferentes variantes obtenemos los siguientes ordenamientos:

**Para el Query 3**

| Salida MATLAB-Escalar | | | Salida MATLAB-Coseno | |
|---|---|---|---|---|
| Document 5 - Similarity: 11 | | | Document 5 - Similarity: 0.52204 | |
| Document 6 - Similarity: 4 | | | Document 6 - Similarity: 0.20966 | |
| Escalar + Stemm | | | Coseno+Stemm | |
| Document 5 - Similarity: 15 | | | Document 5 - Similarity: 0.56898 | |
| Document 6 - Similarity: 8 | | | Document 6 - Similarity: 0.36326 | |
| Document 2 - Similarity: 4 | | | Document 2 - Similarity: 0.09673 | |
| Document 8 - Similarity: 1 | | | Document 4 - Similarity: 0.035578 | |
| Document 4 - Similarity: 1 | | | Document 1 - Similarity: 0.035466 | |
| Document 1 - Similarity: 1 | | | Document 8 - Similarity: 0.027682 | |
| Escalar + Stemm + IDF | | | Coseno + Stemm + IDF | |
| Document 5 - Similarity: 35.8289 | | | Document 5 - Similarity: 0.57527 | |
| Document 6 - Similarity: 17.5754 | | | Document 6 - Similarity: 0.32278 | |
| Document 2 - Similarity: 5.2877 | | | Document 2 - Similarity: 0.065922 | |
| Document 8 - Similarity: 2.3219 | | | Document 4 - Similarity: 0.038758 | |
| Document 4 - Similarity: 1.3219 | | | Document 8 - Similarity: 0.034669 | |
| Document 1 - Similarity: 1.3219 | | | Document 1 - Similarity: 0.015348 | |

Aquí se observa practicamente el mismo ordenamiento para ambas funciones y variantes presentándose solo una diferencia entre el documento 8 y 1 aplicando coseno con ponderación simple y coseno con ponderación IDF.

**Para el Query 65**

| Salida MATLAB_Escalar | | | Salida MATLAB-Coseno | |
|---|---|---|---|---|
| Document 2 - Similarity: 28 | | | Document 3 - Similarity: 0.76332 | |
| Document 9 - Similarity: 25 | | | Document 4 - Similarity: 0.64693 | |
| Document 4 - Similarity: 23 | | | Document 2 - Similarity: 0.53846 | |
| Document 3 - Similarity: 17 | | | Document 9 - Similarity: 0.39371 | |
| Document 7 - Similarity: 13 | | | Document 7 - Similarity: 0.24533 | |
| Document 8 - Similarity: 8 | | | Document 8 - Similarity: 0.18373 | |
| Document 6 - Similarity: 3 | | | Document 6 - Similarity: 0.11119 | |
| Document 10 - Similarity: 2 | | | Document 10 - Similarity: 0.084515 | |
| Escalar + Stemm | | | Coseno + Stemm | |
| Document 2 - Similarity: 28 | | | Document 3 - Similarity: 0.71967 | |
| Document 9 - Similarity: 25 | | | Document 4 - Similarity: 0.60993 | |
| Document 4 - Similarity: 23 | | | Document 2 - Similarity: 0.50469 | |

| | | |
|---|---|---|
| Document 3 - Similarity: 17 | | Document 9 - Similarity: 0.3509 |
| Document 7 - Similarity: 14 | | Document 7 - Similarity: 0.24099 |
| Document 8 - Similarity: 8 | | Document 8 - Similarity: 0.16506 |
| Document 10 - Similarity: 4 | | Document 10 - Similarity: 0.14907 |
| Document 6 - Similarity: 3 | | Document 6 - Similarity: 0.10153 |
| Document 1 - Similarity: 3 | | Document 1 - Similarity: 0.079305 |
| Document 5 - Similarity: 2 | | Document 5 - Similarity: 0.056546 |
| Escalar + Stemm + IDF | | Coseno + Stemm + IDF |
| Document 2 - Similarity: 18.9007 | | Document 3 - Similarity: 0.38655 |
| Document 9 - Similarity: 15.8651 | | Document 4 - Similarity: 0.30646 |
| Document 3 - Similarity: 15.8312 | | Document 2 - Similarity: 0.17563 |
| Document 4 - Similarity: 14.0235 | | Document 10 - Similarity: 0.12823 |
| Document 10 - Similarity: 9.2877 | | Document 9 - Similarity: 0.11729 |
| Document 7 - Similarity: 8.1985 | | Document 7 - Similarity: 0.068266 |
| Document 8 - Similarity: 3.9997 | | Document 8 - Similarity: 0.044513 |
| Document 6 - Similarity: 1.1811 | | Document 6 - Similarity: 0.016169 |
| Document 1 - Similarity: 1.1811 | | Document 1 - Similarity: 0.010222 |
| Document 5 - Similarity: 0.30401 | | Document 5 - Similarity: 0.0036382 |

Aquí se observa como detalle más significativo, que el Documento Nro. 10 resulta beneficiado en el ordenamiento con ambas funciones al aplicar ponderación IDF. Esto se debe seguramente a que alguno de los términos que posee este documento son más raros dentro de la colección de los 10 documentos y esto lo ayudaría a subir en el ranking general.

**Para el Query 67**

| | | |
|---|---|---|
| Salida MATLAB-Euclideas | | Salida MATLAB-Coseno |
| Document 9 - Similarity: 28 | | Document 3 - Similarity: 0.72002 |
| Document 2 - Similarity: 28 | | Document 4 - Similarity: 0.69159 |
| Document 4 - Similarity: 23 | | Document 2 - Similarity: 0.57564 |
| Document 3 - Similarity: 15 | | Document 9 - Similarity: 0.4714 |
| Document 7 - Similarity: 13 | | Document 7 - Similarity: 0.26227 |
| Document 8 - Similarity: 9 | | Document 8 - Similarity: 0.22096 |
| Document 1 - Similarity: 4 | | Document 1 - Similarity: 0.12066 |
| Document 6 - Similarity: 3 | | Document 6 - Similarity: 0.11886 |
| Document 5 - Similarity: 2 | | Document 5 - Similarity: 0.07175 |
| Euclideas + Stemm | | Coseno + Stemm |
| Document 9 - Similarity: 28 | | Document 3 - Similarity: 0.72002 |
| Document 2 - Similarity: 28 | | Document 4 - Similarity: 0.69159 |
| Document 4 - Similarity: 23 | | Document 2 - Similarity: 0.57226 |
| Document 3 - Similarity: 15 | | Document 9 - Similarity: 0.44562 |
| Document 7 - Similarity: 13 | | Document 7 - Similarity: 0.25373 |
| Document 8 - Similarity: 9 | | Document 8 - Similarity: 0.21056 |
| Document 1 - Similarity: 4 | | Document 1 - Similarity: 0.1199 |
| Document 6 - Similarity: 3 | | Document 6 - Similarity: 0.11513 |
| Document 5 - Similarity: 2 | | Document 5 - Similarity: 0.064117 |
| Euclideas + Stemm + IDF | | Coseno + Stemm + IDF |
| Document 9 - Similarity: 21.076 | | Document 4 - Similarity: 0.3475 |

| | | |
|---|---|---|
| Document 2 - Similarity: 18.9007 | | Document 3 - Similarity: 0.28205 |
| Document 4 - Similarity: 14.0235 | | Document 2 - Similarity: 0.19915 |
| Document 3 - Similarity: 10.1873 | | Document 9 - Similarity: 0.17668 |
| Document 8 - Similarity: 5.7367 | | Document 8 - Similarity: 0.072392 |
| Document 7 - Similarity: 4.8766 | | Document 7 - Similarity: 0.046042 |
| Document 1 - Similarity: 2.9181 | | Document 1 - Similarity: 0.028634 |
| Document 6 - Similarity: 1.1811 | | Document 6 - Similarity: 0.018333 |
| Document 5 - Similarity: 0.30401 | | Document 5 - Similarity: 0.0041253 |

Para esta interrogación observamos la diferencia que ya se mencionó anteriormente sobre el ordenamiento de los primeros documentos del ranking.

# Capítulo 10

## Estudio 3: Análisis de Semántica Latente

### Objetivo

El objetivo de este estudio es aplicar sobre algunos registros de la colección de prueba Cranfield el tratamiento necesario para:

1) Mostrar de forma comparativa los resultados obtenidos al contrastar las 10 interrogaciones elegidas contra la representación de los resúmenes de 10 documentos realizada según:
   a. Espacio vectorial
   b. Semántica Latente seleccionando diferente cantidad de factores
2) Mostrar el espacio documental representado gráficamente a partir de los títulos de dos manera distintas
   a. Matriz término/documento aplicando Escalamiento Multidimensional (MDS)
   b. Matrices resultantes de aplicar SVD usando un gráfico BIPLOT.

### Fuente de datos

Los datos utilizados en este estudio corresponden a la colección Cranfield ya descrita en el capítulo anterior (ver p.103).

### Procesamiento de los textos

Se realizó sobre los registros seleccionados de la colección una serie de tareas utilizando diferentes herramientas computacionales con la finalidad de obtener los resultados propuestos en los objetivos. Se resume a continuación la secuencia seguida:

**Paso 1**: Aplicando el utilitario MX (CISIS) se generó a partir del archivo de texto de la colección original una base de datos Isis que contendrá los 1400 registros completos.

Luego se extrajo el campo correspondiente al resumen y se generó un nuevo archivo de texto[1] que servirá como insumo en el paso siguiente.

**Paso 2**: Utilizando la librería desarrollada en MatLab ya mencionada en el estudio anterior, se obtuvo la matriz término/documento de la totalidad de los resúmenes de la colección y se aplicó el Análisis de Semántica Latente tomando diferente número de factores[2].

**Paso 3**: Utilizando el módulo de Recuperación de Información de la misma libreria se contrastaron las 10 primeras interrogaciones de la colección Cranfield contra las diferentes matrices obtenidas en el paso anterior. y con la matriz generada por el método del Espacio Vectorial. El cálculo de similitud se realizó aplicando la función del coseno[3].

**Paso 4**: Se analizaron los resultados de la siguiente manera:

a) Por cada interrogación se tomaron solamente los 5 primeros resultados obtenidos
b) Se colorearon de acuerdo con los juicios de relevancia de la colección: rojo para el nivel 1, azul para el nivel 2, verde para el nivel 3 y violeta para el nivel 4
c) Se generó un indicador para ponderar los resultados teniendo en cuenta las variaciones del coeficiente de similitud, el juicio de relevancia y el orden en la salida[4]:

**coef.simil. * (1/valor juicio relev.) * valor orden de 5 a 1**

d) Se elaboró una tabla comparativa con los indicadores y se obtuvieron los promedios

**Paso 5**: Se generó una matriz término/documento para los 10 primeros títulos de la colección y se aplicó Semántica Latente. Se graficó el resultado utilizando una macro de EXCEL para generar gráficos tipo BIPLOT.

[1] Archivo ***cran_1400_Abs.txt*** disponibles en el CD en \Tesina\Experimental\Estudio 3
[2] Directorio ***cran_Abs_1400*** disponible en el CD en \Tesina\Experimental\Estudio 3
[3] Archivo ***cran_Abs_1400.xsl*** disponible en el CD en \Tesina\Experimental\Estudio 3
[4] Cuando dos resultados corresponden al mismo nivel (mismo color) y están adyacentes, se aplica el mismo factor multiplicador de orden.

**Paso 6**: Utilizando la matriz término/documento del paso anterior se graficó utilizando el módulo MDS (Escalamiento Multidimensional) del paquete estadístico SPSS.

## Resultados

Habiendo aplicado los procesos antes mencionados sobre la colección de los 10 registros bibliográficos que se detallaron en el Estudio 2 (ver p.105) y utilizando las 10 interrogaciones siguientes[5], se han obtenido los siguientes resultados:

Query 1

.W
what similarity laws must be obeyed when constructing aeroelastic models of heated high speed aircraft .

| ESPACIO VECTORIAL |
|---|
| **Document 51 - Similarity: 0.43036** |
| **Document 12 - Similarity: 0.36479** |
| **Document 486 - Similarity: 0.33896** |
| **Document 184 - Similarity: 0.29488** |
| Document 877 - Similarity: 0.2883 |
| Indicador: 2,9319 |

| SEMANTICA LATENTE (2 factores) | SEMANTICA LATENTE (10 factores) |
|---|---|
| Document 840 - Similarity: 0.12059 | Document 707 - Similarity: 0.27765 |
| Document 903 - Similarity: 0.12059 | **Document 102 - Similarity: 0.27129** |
| Document 829 - Similarity: 0.12059 | Document 1099 - Similarity: 0.26865 |
| Document 617 - Similarity: 0.12059 | Document 437 - Similarity: 0.26575 |
| **Document 15 - Similarity: 0.12059** | Document 436 - Similarity: 0.26546 |
| Indicador: 0,1507 | Indicador: 0,2719 |

| SEMANTICA LATENTE (30 factores) | SEMANTICA LATENTE (100 factores) |
|---|---|
| **Document 12 - Similarity: 0.3339** | **Document 51 - Similarity: 0.40863** |
| Document 436 - Similarity: 0.31749 | **Document 12 - Similarity: 0.39023** |
| **Document 102 - Similarity: 0.30793** | **Document 486 - Similarity: 0.37498** |
| **Document 51 - Similarity: 0.30688** | **Document 102 - Similarity: 0.33999** |
| Document 429 - Similarity: 0.29542 | Document 874 - Similarity: 0.32791 |
| Indicador: 1,1713 | Indicador: 2,6830 |

| SEMANTICA LATENTE (200 factores) | SEMANTICA LATENTE (300 factores) |
|---|---|
| **Document 51 - Similarity: 0.42957** | **Document 51 - Similarity: 0.4329** |
| **Document 486 - Similarity: 0.419** | **Document 486 - Similarity: 0.39436** |
| **Document 12 - Similarity: 0.37438** | **Document 12 - Similarity: 0.36685** |
| **Document 184 - Similarity: 0.32712** | **Document 102 - Similarity: 0.30384** |
| **Document 102 - Similarity: 0.32318** | **Document 184 - Similarity: 0.29077** |
| Indicador: 3,2001 | Indicador: 3,1150 |

[5] En el archivo original “cran.qry” las interrogaciones no se encuentran numeradas de manera correlativa, aunque si lo están en el archivo que contiene los juicios de relevancia. Se prefirió aquí reenumerarlas de manera consecutiva tal como lo están en el archivo “cranqrel”.

Query 2

.W

what are the structural and aeroelastic problems associated with flight of high speed aircraft .

| ESPACIO VECTORIAL |
|---|
| **Document 12 - Similarity: 0.70432** |
| **Document 746 - Similarity: 0.38778** |
| Document 1169 - Similarity: 0.37796 |
| **Document 51 - Similarity: 0.37589** |
| Document 92 - Similarity: 0.35921 |
| Indicador: 5,7111 |

| SEMANTICA LATENTE (2 factores) | SEMANTICA LATENTE (10 factores) |
|---|---|
| Document 739 - Similarity: 0.098518 | **Document 12 - Similarity: 0.18883** |
| Document 301 - Similarity: 0.098518 | Document 543 - Similarity: 0.18556 |
| Document 949 - Similarity: 0.098517 | Document 834 - Similarity: 0.18461 |
| Document 66 - Similarity: 0.098517 | Document 156 - Similarity: 0.18415 |
| Document 566 - Similarity: 0.098517 | Document 462 - Similarity: 0.18359 |
| Indicador: 0 | Indicador: 0,9442 |

| SEMANTICA LATENTE (30 factores) | SEMANTICA LATENTE (100 factores) |
|---|---|
| **Document 12 - Similarity: 0.44797** | **Document 12 - Similarity: 0.65075** |
| Document 429 - Similarity: 0.42146 | Document 92 - Similarity: 0.53178 |
| Document 92 - Similarity: 0.41943 | Document 1169 - Similarity: 0.45072 |
| Document 578 - Similarity: 0.40019 | Document 429 - Similarity: 0.44887 |
| Document 588 - Similarity: 0.40008 | **Document 746 - Similarity: 0.41481** |
| Indicador: 2,2399 | Indicador: 3,6686 |

| SEMANTICA LATENTE (200 factores) | SEMANTICA LATENTE (300 factores) |
|---|---|
| **Document 12 - Similarity: 0.69745** | **Document 12 - Similarity: 0.71024** |
| Document 92 - Similarity: 0.44835 | Document 92 - Similarity: 0.4241 |
| Document 429 - Similarity: 0.42739 | Document 1169 - Similarity: 0.4186 |
| Document 1169 - Similarity: 0.4272 | Document 429 - Similarity: 0.41841 |
| **Document 746 - Similarity: 0.39351** | **Document 51 - Similarity: 0.38865** |
| Indicador: 3,8808 | Indicador: 1,0989 |

Query 3

.W

what problems of heat conduction in composite slabs have been solved so far .

| ESPACIO VECTORIAL |
|---|
| **Document 5 - Similarity: 0.53675** |
| **Document 485 - Similarity: 0.4901** |
| **Document 181 - Similarity: 0.46981** |
| **Document 399 - Similarity: 0.40269** |

| Document 144 - Similarity: 0.39441 |
|---|
| Indicador: 3,1417 |

| SEMANTICA LATENTE (2 factores) | SEMANTICA LATENTE (10 factores) |
|---|---|
| Document 21 - Similarity: 0.13573 | Document 509 - Similarity: 0.31808 |
| Document 664 - Similarity: 0.13573 | Document 982 - Similarity: 0.3162 |
| **Document 6 - Similarity: 0.13573** | Document 158 - Similarity: 0.31469 |
| Document 1241 - Similarity: 0.13573 | **Document 5 - Similarity: 0.31345** |
| Document 387 - Similarity: 0.13573 | Document 396 - Similarity: 0.31251 |
| Indicador: 0,1357 | Indicador: 0,3135 |

| SEMANTICA LATENTE (30 factores) | SEMANTICA LATENTE (100 factores) |
|---|---|
| Document 396 - Similarity: 0.37381 | **Document 181 - Similarity: 0.43451** |
| **Document 5 - Similarity: 0.37165** | Document 542 - Similarity: 0.41185 |
| Document 51 - Similarity: 0.36008 | **Document 5 - Similarity: 0.3969** |
| Document 542 - Similarity: 0.34367 | Document 587 - Similarity: 0.36672 |
| **Document 181 - Similarity: 0.34127** | **Document 399 - Similarity: 0.36215** |
| Indicador: 0,7129 | Indicador: 1,2418 |

| SEMANTICA LATENTE (200 factores) | SEMANTICA LATENTE (300 factores) |
|---|---|
| **Document 181 - Similarity: 0.48194** | **Document 181 - Similarity: 0.51312** |
| **Document 5 - Similarity: 0.42895** | **Document 5 - Similarity: 0.48548** |
| **Document 399 - Similarity: 0.38088** | **Document 399 - Similarity: 0.40163** |
| Document 542 - Similarity: 0.37747 | Document 542 - Similarity: 0.35847 |
| Document 585 - Similarity: 0.36867 | Document 509 - Similarity: 0.35807 |
| Indicador: 2,1530 | Indicador: 2,3337 |

Query 4

.W

can a criterion be developed to show empirically the validity of flow solutions for chemically reacting gas mixtures based on the simplifying assumption of instantaneous local chemical equilibrium .

| ESPACIO VECTORIAL |
|---|
| Document 167 - Similarity: 0.31693 |
| Document 166 - Similarity: 0.29753 |
| Document 1085 - Similarity: 0.25649 |
| Document 188 - Similarity: 0.25541 |
| Document 575 - Similarity: 0.25058 |
| Indicador: 0 |

| SEMANTICA LATENTE (2 factores) | SEMANTICA LATENTE (10 factores) |
|---|---|
| Document 500 - Similarity: 0.1872 | Document 1194 - Similarity: 0.28175 |
| Document 1288 - Similarity: 0.1872 | Document 660 - Similarity: 0.27953 |
| Document 565 - Similarity: 0.1872 | Document 167 - Similarity: 0.27907 |
| Document 573 - Similarity: 0.1872 | Document 989 - Similarity: 0.27841 |

| | |
|---|---|
| Document 128 - Similarity: 0.1872 | **Document 236 - Similarity: 0.27793** |
| Indicador: 0 | Indicador: 0,0926 |

| **SEMANTICA LATENTE (30 factores)** | **SEMANTICA LATENTE (100 factores)** |
|---|---|
| Document 167 - Similarity: 0.31453 | Document 167 - Similarity: 0.35867 |
| Document 1245 - Similarity: 0.29194 | **Document 236 - Similarity: 0.32139** |
| **Document 236 - Similarity: 0.28824** | Document 317 - Similarity: 0.31034 |
| Document 1189 - Similarity: 0.28389 | Document 541 - Similarity: 0.30552 |
| Document 118 - Similarity: 0.28374 | Document 1189 - Similarity: 0.30017 |
| Indicador: 0,2882 | Indicador: 0,4285 |

| **SEMANTICA LATENTE (200 factores)** | **SEMANTICA LATENTE (300 factores)** |
|---|---|
| Document 167 - Similarity: 0.35146 | Document 167 - Similarity: 0.33982 |
| Document 317 - Similarity: 0.3018 | **Document 236 - Similarity: 0.29685** |
| **Document 236 - Similarity: 0.29632** | **Document 166 - Similarity: 0.29536** |
| **Document 166 - Similarity: 0.29247** | Document 28 - Similarity: 0.26618 |
| Document 28 - Similarity: 0.28035 | Document 317 - Similarity: 0.26121 |
| Indicador: 0,5888 | Indicador: 0,7896 |

Query 5

.W

what chemical kinetic system is applicable to hypersonic aerodynamic problems .

| **ESPACIO VECTORIAL** |
|---|
| Document 1379 - Similarity: 0.28697 |
| Document 368 - Similarity: 0.25928 |
| Document 367 - Similarity: 0.2582 |
| Document 641 - Similarity: 0.25198 |
| Document 1032 - Similarity: 0.2357 |
| Indicador: 0 |

| **SEMANTICA LATENTE (2 factores)** | **SEMANTICA LATENTE (10 factores)** |
|---|---|
| Document 356 - Similarity: 0.088726 | Document 619 - Similarity: 0.14234 |
| Document 164 - Similarity: 0.088726 | Document 617 - Similarity: 0.14205 |
| Document 668 - Similarity: 0.088726 | Document 614 - Similarity: 0.14054 |
| Document 1309 - Similarity: 0.088726 | Document 613 - Similarity: 0.13983 |
| Document 332 - Similarity: 0.08872 | Document 1298 - Similarity: 0.13958 |
| Indicador: 0 | Indicador: 0 |

| **SEMANTICA LATENTE (30 factores)** | **SEMANTICA LATENTE (100 factores)** |
|---|---|
| Document 1379 - Similarity: 0.27303 | Document 1379 - Similarity: 0.36171 |
| Document 368 - Similarity: 0.25302 | Document 368 - Similarity: 0.32865 |
| Document 281 - Similarity: 0.25264 | Document 746 - Similarity: 0.31128 |
| Document 92 - Similarity: 0.25139 | Document 641 - Similarity: 0.29295 |
| Document 778 - Similarity: 0.24319 | Document 1253 - Similarity: 0.25504 |
| Indicador: 0 | Indicador: 0 |

| SEMANTICA LATENTE (200 factores) | SEMANTICA LATENTE (300 factores) |
|---|---|
| Document 1379 - Similarity: 0.3397 | Document 1379 - Similarity: 0.32991 |
| Document 368 - Similarity: 0.3001 | Document 368 - Similarity: 0.29668 |
| Document 746 - Similarity: 0.29273 | Document 367 - Similarity: 0.27385 |
| Document 641 - Similarity: 0.28453 | Document 746 - Similarity: 0.27366 |
| Document 730 - Similarity: 0.24138 | Document 641 - Similarity: 0.27227 |
| Indicador: 0 | Indicador: 0 |

Query 6

.W

what theoretical and experimental guides do we have as to turbulent couette flow behaviour .

| ESPACIO VECTORIAL |
|---|
| **Document 491 - Similarity: 0.40731** |
| **Document 257 - Similarity: 0.314** |
| Document 1275 - Similarity: 0.28962 |
| Document 386 - Similarity: 0.28697 |
| Document 775 - Similarity: 0.27954 |
| Indicador: 2,4552 |

| SEMANTICA LATENTE (2 factores) | SEMANTICA LATENTE (10 factores) |
|---|---|
| Document 131 - Similarity: 0.20906 | Document 1275 - Similarity: 0.35579 |
| Document 1194 - Similarity: 0.20906 | Document 751 - Similarity: 0.35399 |
| Document 1050 - Similarity: 0.20906 | Document 892 - Similarity: 0.3513 |
| Document 935 - Similarity: 0.20906 | Document 900 - Similarity: 0.35051 |
| Document 767 - Similarity: 0.20906 | Document 18 - Similarity: 0.34867 |
| Indicador: 0 | Indicador:0 |

| SEMANTICA LATENTE (30 factores) | SEMANTICA LATENTE (100 factores) |
|---|---|
| Document 1275 - Similarity: 0.40311 | Document 1275 - Similarity: 0.41288 |
| Document 18 - Similarity: 0.3634 | **Document 491 - Similarity: 0.37869** |
| **Document 491 - Similarity: 0.35825** | **Document 257 - Similarity: 0.34682** |
| Document 418 - Similarity: 0.35491 | Document 17 - Similarity: 0.34064 |
| Document 427 - Similarity: 0.35249 | Document 1374 - Similarity: 0.33022 |
| Indicador: 1,0748 | Indicador: 0,7255 |

| SEMANTICA LATENTE (200 factores) | SEMANTICA LATENTE (300 factores) |
|---|---|
| **Document 491 - Similarity: 0.38098** | **Document 491 - Similarity: 0.36611** |
| Document 1275 - Similarity: 0.3505 | Document 1275 - Similarity: 0.32346 |
| **Document 257 - Similarity: 0.33148** | **Document 257 - Similarity: 0.32029** |
| Document 1374 - Similarity: 0.32756 | Document 1374 - Similarity: 0.30818 |
| **Document 258 - Similarity: 0.32273** | **Document 258 - Similarity: 0.30627** |
| Indicador: 2,3440 | Indicador: 2,2529 |

Query 7

.W

is it possible to relate the available pressure distributions for an ogive forebody at zero angle of attack to the lower surface pressures of an equivalent ogive forebody at angle of attack .

| ESPACIO VECTORIAL |
|---|
| **Document 492 - Similarity: 0.81524** |
| Document 1231 - Similarity: 0.47756 |
| Document 1307 - Similarity: 0.42787 |
| Document 354 - Similarity: 0.41144 |
| Document 197 - Similarity: 0.35001 |
| Indicador: 4,0762 |

| SEMANTICA LATENTE (2 factores) | SEMANTICA LATENTE (10 factores) |
|---|---|
| Document 433 - Similarity: 0.28434 | **Document 492 - Similarity: 0.35656** |
| Document 747 - Similarity: 0.28434 | Document 994 - Similarity: 0.35413 |
| Document 696 - Similarity: 0.28434 | Document 605 - Similarity: 0.3533 |
| Document 1164 - Similarity: 0.28434 | Document 1352 - Similarity: 0.34276 |
| Document 1001 - Similarity: 0.28434 | Document 1350 - Similarity: 0.34055 |
| Indicador: 0 | Indicador: 1,7828 |

| SEMANTICA LATENTE (30 factores) | SEMANTICA LATENTE (100 factores) |
|---|---|
| **Document 492 - Similarity: 0.58401** | **Document 492 - Similarity: 0.64427** |
| Document 1231 - Similarity: 0.51661 | Document 1231 - Similarity: 0.54073 |
| Document 423 - Similarity: 0.51528 | Document 1307 - Similarity: 0.45162 |
| Document 1285 - Similarity: 0.49262 | Document 354 - Similarity: 0.41865 |
| Document 1307 - Similarity: 0.48222 | Document 1285 - Similarity: 0.41266 |
| Indicador: 2,9201 | Indicador: 3,2214 |

| SEMANTICA LATENTE (200 factores) | SEMANTICA LATENTE (300 factores) |
|---|---|
| **Document 492 - Similarity: 0.66636** | **Document 492 - Similarity: 0.676** |
| Document 1231 - Similarity: 0.51956 | Document 1231 - Similarity: 0.51063 |
| Document 1307 - Similarity: 0.44653 | Document 1307 - Similarity: 0.44381 |
| Document 354 - Similarity: 0.43163 | Document 354 - Similarity: 0.42868 |
| **Document 58 - Similarity: 0.38498** | Document 197 - Similarity: 0.37665 |
| Indicador: 3,4601 | Indicador: 3,3800 |

Query 8

.W

what methods -dash exact or approximate -dash are presently available for predicting body pressures at angle of attack.

| ESPACIO VECTORIAL |
|---|
| **Document 492 - Similarity: 0.56527** |
| **Document 122 - Similarity: 0.45441** |
| Document 1231 - Similarity: 0.42144 |
| Document 234 - Similarity: 0.39693 |
| **Document 354 - Similarity: 0.3473** |

Indicador: 5,4457

| SEMANTICA LATENTE (2 factores) | SEMANTICA LATENTE (10 factores) |
|---|---|
| Document 1247 - Similarity: 0.29332 | **Document 492 - Similarity: 0.42876** |
| Document 1066 - Similarity: 0.29332 | **Document 122 - Similarity: 0.42549** |
| Document 807 - Similarity: 0.29332 | **Document 1005 - Similarity: 0.41554** |
| Document 442 - Similarity: 0.29332 | Document 234 - Similarity: 0.41467 |
| Document 605 - Similarity: 0.29332 | Document 248 - Similarity: 0.4143 |
| Indicador: 0 | Indicador: 6,3490 |

| SEMANTICA LATENTE (30 factores) | SEMANTICA LATENTE (100 factores) |
|---|---|
| **Document 122 - Similarity: 0.4818** | **Document 492 - Similarity: 0.54603** |
| **Document 492 - Similarity: 0.47395** | **Document 122 - Similarity: 0.50018** |
| Document 234 - Similarity: 0.4554 | Document 1231 - Similarity: 0.44374 |
| Document 1262 - Similarity: 0.43498 | Document 473 - Similarity: 0.40892 |
| Document 248 - Similarity: 0.43074 | Document 1262 - Similarity: 0.39617 |
| Indicador: 4,7788 | Indicador: 5,2311 |

| SEMANTICA LATENTE (200 factores) | SEMANTICA LATENTE (300 factores) |
|---|---|
| **Document 492 - Similarity: 0.58498** | **Document 492 - Similarity: 0.59621** |
| **Document 122 - Similarity: 0.46735** | **Document 122 - Similarity: 0.46829** |
| Document 1231 - Similarity: 0.45006 | Document 1231 - Similarity: 0.44263 |
| Document 234 - Similarity: 0.39507 | Document 234 - Similarity: 0.40025 |
| Document 473 - Similarity: 0.37744 | Document 473 - Similarity: 0.36772 |
| Indicador: 5,2617 | Indicador: 5,3225 |

Query 9

```
.W
papers on internal /slip flow/ heat transfer studies .
```

| ESPACIO VECTORIAL |
|---|
| Document 398 - Similarity: 0.48107 |
| Document 45 - Similarity: 0.47173 |
| **Document 550 - Similarity: 0.44882** |
| Document 983 - Similarity: 0.42521 |
| **Document 21 - Similarity: 0.4234** |

Indicador: 0,8849

| SEMANTICA LATENTE (2 factores) | SEMANTICA LATENTE (10 factores) |
|---|---|
| Document 325 - Similarity: 0.28592 | Document 407 - Similarity: 0.50971 |
| Document 1154 - Similarity: 0.28592 | **Document 550 - Similarity: 0.50713** |
| Document 1244 - Similarity: 0.28591 | Document 387 - Similarity: 0.50328 |
| Document 386 - Similarity: 0.28591 | Document 269 - Similarity: 0.49907 |
| Document 582 - Similarity: 0.28591 | Document 500 - Similarity: 0.49736 |
| Indicador: 0 | Indicador: 1,0143 |

| SEMANTICA LATENTE (30 factores) | SEMANTICA LATENTE (100 factores) |
|---|---|
| Document 398 - Similarity: 0.54382 | Document 398 - Similarity: 0.52068 |
| Document 873 - Similarity: 0.53238 | Document 873 - Similarity: 0.50497 |
| Document 102 - Similarity: 0.49344 | Document 872 - Similarity: 0.47575 |
| Document 396 - Similarity: 0.48907 | Document 623 - Similarity: 0.46136 |
| Document 500 - Similarity: 0.48873 | Document 303 - Similarity: 0.45862 |
| Indicador: 0 | Indicador: 0 |

| SEMANTICA LATENTE (200 factores) | SEMANTICA LATENTE (300 factores) |
|---|---|
| Document 398 - Similarity: 0.5394 | Document 398 - Similarity: 0.53527 |
| Document 872 - Similarity: 0.46996 | Document 45 - Similarity: 0.46404 |
| Document 303 - Similarity: 0.46807 | Document 983 - Similarity: 0.46052 |
| Document 983 - Similarity: 0.46239 | Document 872 - Similarity: 0.4457 |
| Document 45 - Similarity: 0.44777 | Document 303 - Similarity: 0.44125 |
| Indicador: 0 | Indicador: 0 |

Query 10

.W

are real-gas transport properties for air available over a wide range of enthalpies and densities .

| ESPACIO VECTORIAL |
|---|
| **Document 302 - Similarity: 0.38231** |
| **Document 493 - Similarity: 0.32532** |
| Document 1143 - Similarity: 0.29839 |
| Document 949 - Similarity: 0.28897 |
| Document 616 - Similarity: 0.27854 |
| Indicador: 1,9385 |

| SEMANTICA LATENTE (2 factores) | SEMANTICA LATENTE (10 factores) |
|---|---|
| Document 596 - Similarity: 0.075646 | Document 552 - Similarity: 0.14335 |
| Document 76 - Similarity: 0.075646 | Document 865 - Similarity: 0.14161 |
| Document 770 - Similarity: 0.075646 | Document 1318 - Similarity: 0.13872 |
| Document 848 - Similarity: 0.075646 | Document 169 - Similarity: 0.13793 |
| Document 1248 - Similarity: 0.075646 | Document 1158 - Similarity: 0.13653 |
| Indicador: 0 | Indicador: 0 |

| SEMANTICA LATENTE (30 factores) | SEMANTICA LATENTE (100 factores) |
|---|---|
| **Document 405 - Similarity: 0.27207** | **Document 405 - Similarity: 0.37791** |
| Document 488 - Similarity: 0.25477 | **Document 302 - Similarity: 0.36686** |
| Document 185 - Similarity: 0.25004 | Document 691 - Similarity: 0.31186 |
| Document 1286 - Similarity: 0.24611 | Document 185 - Similarity: 0.30489 |
| **Document 1011 - Similarity: 0.24365** | **Document 1011 - Similarity: 0.28767** |
| Indicador: 0,7614 | Indicador: 1,5298 |

| SEMANTICA LATENTE (200 factores) | SEMANTICA LATENTE (300 factores) |
|---|---|
| **Document 405 - Similarity: 0.38018** | **Document 405 - Similarity: 0.38045** |
| **Document 302 - Similarity: 0.35962** | **Document 302 - Similarity: 0.37657** |
| Document 185 - Similarity: 0.3173 | **Document 493 - Similarity: 0.30551** |
| Document 691 - Similarity: 0.29779 | Document 616 - Similarity: 0.29614 |
| **Document 437 - Similarity: 0.29152** | Document 949 - Similarity: 0.28953 |
| Indicador: 1,5271 | Indicador: 2,3697 |

La comparación entre los promedios de los diferentes factores aplicados al Análisis de Semántica Latente arroja que el factor más eficaz en términos de recuperación sería el Factor 200:

| | 2 factores | 10 factores | 30 factores | 100 factores | 200 factores | 300 factores |
|---|---|---|---|---|---|---|
| **Q1** | 0,1507 | 0,2719 | 1,1713 | 2,6830 | 3,2001 | 3,1150 |
| **Q2** | 0,0000 | 0,9442 | 2,2399 | 3,6686 | 3,8808 | 1,0989 |
| **Q3** | 0,1357 | 0,3135 | 0,7129 | 1,2418 | 2,1530 | 2,3337 |
| **Q4** | 0,0000 | 0,0926 | 0,2882 | 0,4285 | 0,5888 | 0,7896 |
| **Q5** | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| **Q6** | 0,0000 | 0,0000 | 1,0748 | 0,7255 | 2,3440 | 2,2529 |
| **Q7** | 0,0000 | 1,7828 | 2,9201 | 3,2214 | 3,4601 | 3,3800 |
| **Q8** | 0,0000 | 6,3490 | 4,7788 | 5,2311 | 5,2617 | 5,3225 |
| **Q9** | 0,0000 | 1,0143 | 0,0000 | 0,0000 | 0,0000 | 0,0000 |
| **Q10** | 0,0000 | 0,0000 | 0,7614 | 1,5298 | 1,5271 | 2,3697 |
| | 0,0286 | 1,0768 | 1,3947 | 1,8730 | 2,2415 | 2,0662 |

Si se compara este valor con el que arroja el promedio obtenido en el método del Espacio Vectorial se observa que para los 10 documentos y 10 interrogaciones seleccionadas, el método del Análisis de Semántica Latente no arroja un resultado superior.

| | Vectorial | 200 factores |
|---|---|---|
| **Q1** | 2,9319 | 3,2001 |
| **Q2** | 5,7111 | 3,8808 |
| **Q3** | 3,1417 | 2,1530 |
| **Q4** | 0,0000 | 0,5888 |
| **Q5** | 0,0000 | 0,0000 |
| **Q6** | 2,4552 | 2,3440 |
| **Q7** | 4,0762 | 3,4601 |
| **Q8** | 5,4457 | 5,2617 |
| **Q9** | 0,8849 | 0,0000 |
| **Q10** | 1,9385 | 1,5271 |
| | 2,6585 | 2,2415 |

Sin embargo, si se analiza particularmente la interrogación 4, se observa que un supuesto SRI que aplicara Semántica Latente estaría mejor preparado para enfrentar el problema de los resultados nulos.

Finalmente, se muestra como se distribuyen los 10 documentos y las palabras de sus títulos luego de aplicar el procesamiento con Análisis de Semántica Latente utilizando un gráfico tipo BIPLOT, señalando, además, la zona de recuperación para la interrogación número 3. Luego se muestra cómo se distribuyen los mismos documentos en el Espacio Vectorial utilizando un gráfico de MDS (Escalamiento Multidimensional) aplicando distancias euclídeas

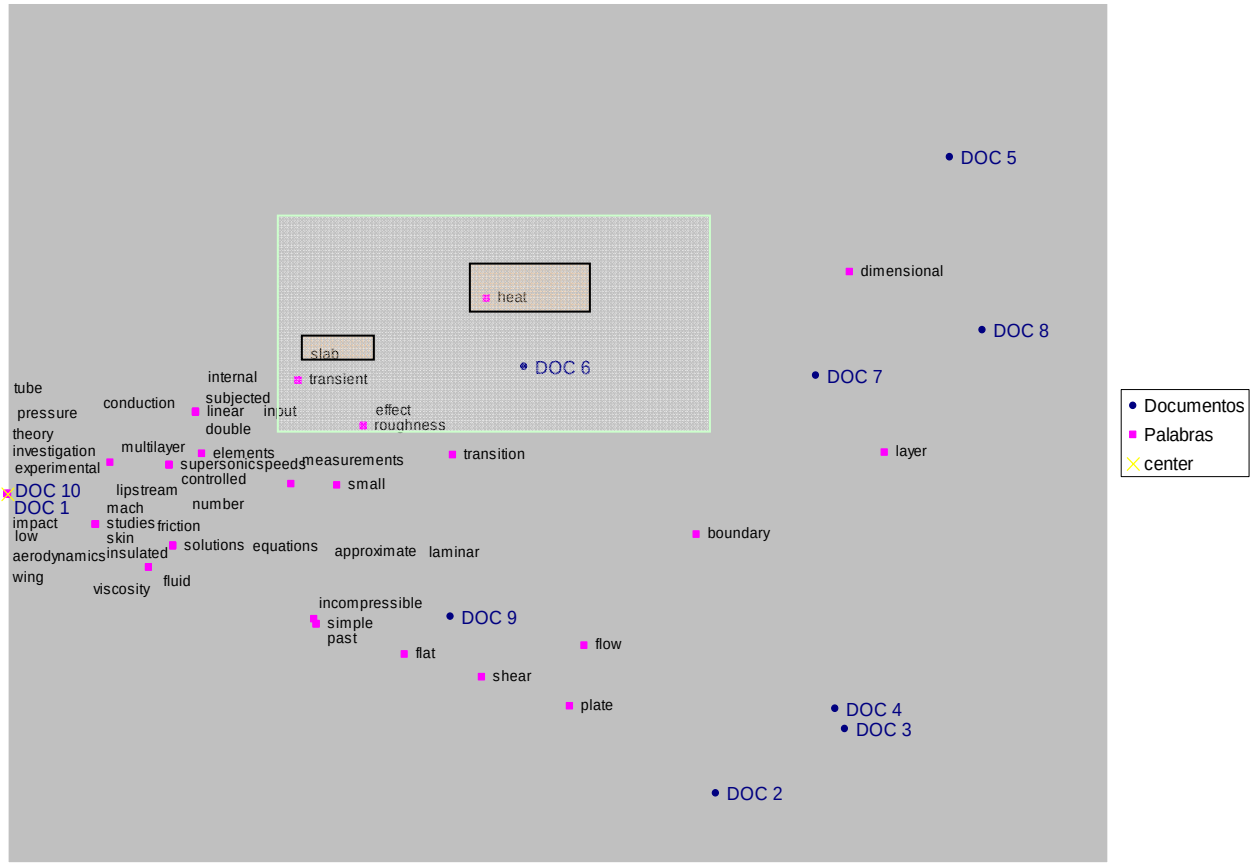



GRAFICO I : Distribución de los 10 primeros títulos de Cranfield aplicando Semántica Latente

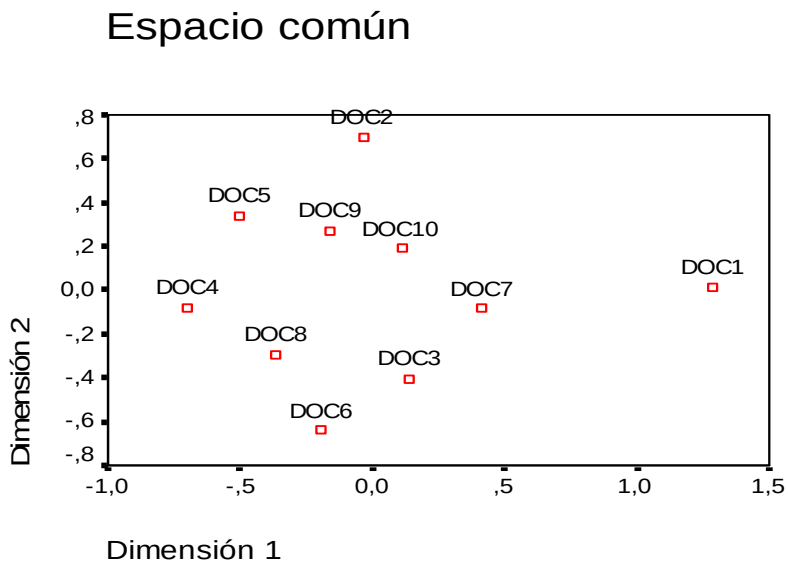



GRAFICO II: Distrib. de los primeros 10 registros de Cranfield aplicando MDS

# Conclusión

Sabemos que el hombre es capaz de determinar el significado de las palabras y las oraciones. Es capaz de realizar juicios sobre la redundancia, la inconsistencia, la relevancia, etc. en los textos y además hacerlo con el lenguaje ordinario, que es extremadamente complejo y poblado de vaguedades y ambigüedades. No existen reglas estrictas sobre qué palabras hay que poner juntas para alcanzar diversos tipos de significados. Frente a este problema, a lo largo de este trabajo se trató de mostrar los procesos básicos que se aplican al tratamiento automático de textos con la finalidad de eliminar vocabulario con poca información, detectar y remover información redundante, extraer información relevante, todo acotado a los Sistemas de Recuperación de Información científico-tecnológica.

El modelo que sustenta estos desarrollos es, en términos generales, un modelo en el que existen objetos documentos cuyas propiedades son las palabras significativas que ellos contienen. Las propiedades son observables y medibles. La medida puede ser simplemente la lista de propiedades que el objeto posee, su frecuencia, su distribución de orden, etc. La información que provee la medida, constituye la evidencia y toda la complejidad de los métodos matemáticos y estadísticos es factible de aplicación en busca de la medida óptima.

Asimismo, el complemento con el conocimiento del área de la lingüística parece necesario y constituye, quizá, el nicho más interesante de trabajo para estas latitudes. Como se ha visto a lo largo de este trabajo, el mundo anglosajón, liderando las investigaciones en RI desde hace años, ha sido prolífico en la realización de estudios, desarrollo de herramientas y adopción de corpus textuales para trabajar con el idioma inglés. El idioma español no ha formado parte significativa de este paradigma de investigación, de la misma manera que no lo han formado los demás idiomas.

Sin embargo, en la actual Sociedad de la Información, donde Internet se ha convertido en la plataforma global de comunicaciones y la Web en el Sistema de Información más heterogéneo, multilingüe y poco estructurado jamás visto, los desafíos para la RI basada en el lenguaje natural y particularmente para el idioma español, parecen renacer. Estudios de revisión como el presente, pero dirigidos exclusivamente al español serían de gran utilidad para complementar el conocimiento de nuestro colectivo profesional sobre la recuperación de información moderna.

Finalmente, se considera que el camino recorrido en la realización de este trabajo es un esfuerzo válido en tanto se convierta en una contribución a la apertura del interés de estudiantes, graduados, docentes e investigadores en un tema fundamental en la

estructura de la Ciencia de la Información. Se espera que el conocimiento sobre el tema evite el direccionamiento de los estudios y esfuerzos de la Ciencia de la Información a la competencia entre hombres y máquinas con las conocidas secuelas de desempleo y conduzca, en cambio, al incremento ilimitado de las capacidades de resolución de problemas y prestación de servicios de los profesionales de la Bibliotecología y Ciencia de la Información, apoyados en la capacidad de uso de las máquinas para la resolución de problemas propios.

# Bibliografía

[1] SPARCK JONES, K. y WILLETT, P. (eds.). *Readings in information retrieval.* San Francisco: Morgan Kaufmann, 1997.

[2] SALTON, G. y McGILL, M.J. *Introduction to modern information retrieval.* New York: McGraw-Hill, 1983.

[3] BAEZA-YATES, R. y RIBEIRO-NETO, B. *Modern information retrieval.* New York: Addison Wesley, 1999.

[4] IGWERSEN, P. *Information retrieval interaction.* Los Angeles: Taylor Graham, 1992.

[5] ELLIS, D. *Progress and problems in information retrieval.* London: Library Association, 1996.

[6] van RIJSBERGEN, C.J. *Information retrieval.* Butterworths, 1979.

[7] BLAIR, D.C. *Language and representation in information retrieval.* Ámsterdam: Elsevier, 1990.

[8] ARCHUBY, C. *Modelos, analogías, metáforas y equivalencias, como instrumentos del trabajo intelectual*. La Plata: UNLP, 2003. [Apunte de cátedra]

[9] BOOKSTEIN, A. Relevance. En: *Journal of the American Society for Information Science*. Sep. 1979, p. 269-273.

[10] PEREZ ALVAREZ-OSSORIO, J. R. *Introducción a la información y documentación científica.* Madrid: Alhambra, 1988.

[11] ROBERTSON, S. E. Theories and models in information retrieval. *Journal of documentation.* 1977, vol.33, p.126-148.

[12] FOSKETT, A.C. *The subject approach to information.* London: Library Association, 1997.

[13] MOREIRO GONZALEZ, J.A. *El contenido de los documentos textuales: su análisis y representación mediante el lenguaje natural.* Gijón: Trea, 2004.

[14] COOPER, W.S. *Getting beyond Boole. Information Processing & Management.* 1988, vol.24, nro.3, p.243-248.

[15] PEÑA, R. *Gestión digital de la información: De bits a bibliotecas digitales y la web.* Madrid: Ra-Ma, 2002.

[16] SALTON, G., WONG, A. y YANG, C.S. A vector space model for automatic indexing. *Communication of the ACM*. 1975, vol. 18, p.613-620.

[17] MOYA ANEGON, Félix de. *Los sistemas integrados de gestión bibliotecaria : Estructura de datos y recuperación de información.* Madrid: ANABAD, 1995.

[18] MARON, M.E. y KUNS, J.L. *On* relevance probabilistic indexing and information retrieval. *Journal of the ACM.* 1960, vol.7, p. 216-244.

[19] ROBERTSON, S.E. The probability ranking principle in IR. *Journal of documentation*. 1977, vol.33, p.294-304.

[20] MOYA ANEGON, F. de, LOPEZ GIJON, J. y GARCIA CARO, C. *Técnicas cuantitativas aplicadas a la biblioteconomía y documentación.* Madrid: Síntesis, 1996.

[21] KRENN, B. y SAMUELSON, C. *The linguist´s guide to statistics: Don´t panic.* [en linea] [consultado Mayo 2003]. Disponible en http://coli.uni-sb.de/

[22] ELRIDGE, R. *Six thousand common english words*. Buffalo: The Clement press, 1911.

[23] ESTOUP, M.J. B. *Gamas estenograficas.* Trad. R. Caballero. Madrid, 1932.

[24] DEWEY, G. *Relative frequency of english speech sounds*. Harvard Univ. Press, 1923.

[25] WYLLYS, R.E. Empirical and theoretical bases of Zipf´s law*. Library Trends.* Summer 1981. p.53-64.

[26] ZIPF, G.K. *The psycho-biology of language: An introduction to dynamic philology.* Cambridge: The MIT Press, 1968.

[27] BROOKES, B.C. y GRIFFITHS, J.M. Frequency-rank distributions. *Journal of the American Society of Information Science.* January, 1978, p.5-13.

[28] BOOTH, A.D. A "law" of occurrences for words of low frequency. *Information and Control.* 1967, vol.10, p.386-393.

[29] BUCKLAND, M.K. y HINDLE, A. Library Zipf*. Journal of Documentation*, 1969, vol.25, no.1, p. 52-56.

[30] EGGHE, L.; ROUSSEAU, R. *Introduction to informetrics : quantitative methods in library , documentation and information science.* Amsterdam: Elsevier, 1990. p.293.

[31] POWERS, D.M.W. Applications and explanations of Zip's Law. *NeMLaP3/CoNLL98: New methods in language processing and computational natural language learning.* ACL, p.151-160.

[32] SUN, Q. ; SHAW, D. y DAVIS, CH. A model for estimating the occurrence of same-frequency words and the boundary between high- and low-frequency words in texts. *Journal of the American Society for Information Science. 1999,* vol.50, no.3, p.280-286.

[33] LI, Wentian. Random texts exhibit Zipf's-Law like word frequency distribution. *IEEE Transactions on information Theory*. 1992*,* vol.38, nro.6, p. 1842-1845.

[34] LI, Wentian. Letters to the editor: comments about "Zipf's law and the structure and evolution of language" of A.A. Tsonis, C. Schultz, P.A. Tsonis. *Complexity.* 1998, vol. 3, nro.5, p.9-10.

[35] EGGHE, L. On the law of Zipf-Mndelbrot for multi-word phrases. *Journal of the American Society for Information Science.* 1999*,* vol.50, nro.3, p.233-241.

[36] MONTEMURRO, M. A. Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Archives on condensed matter.* 2001, vol.2, nro.9 (arXiv:cond-mat/0104066)

[37] LESK, M. The seven ages of information retrieval. UDT Occasional Paper #5. IFLA, 1996.

[38] LUHN, Hans P. The automatic creation of literature abstracts. *IBM Journal.* April 1958, p.159-165.

[39] LUHN, Hans P. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal*. October 1957, p.309-317.

[40] MARON, M.E. Automatic indexing: An experimental inquiry. *Journal of the ACM.* 1961, vol.8, nro.2, p.404-417.

[41] ABRAMSON, N. *Teoría de la Información y codificación*. Madrid: Paraninfo, 1986.

[42] DOYLE, Lauren B. Semantic road maps for literature searchers*. Journal of the ACM*. 1961, vol.8, nro.4, p.553-578.

[43] STILES, H. Edmund. The association factor in information retrieval. *Journal of the ACM*. 1961, vol.8, nro.2, p.271-279.

[44] KROON, F. Linguistic variation in information retrieval using query reformulation. *Thesis of Master of Sciense.* Simon Fraser University, 1999.

[45] HARMAN, D. Automatic indexing. *Interagency Report 4873.* NIST, NLPIR Group Automatic Indexing, 2000.

[46] JAQUEMIN, C. y TZOUKERMANN, E. NLP for term variant extraction: synergy of morphology, lexicon, and syntax. En Strzalkowski, T. (Ed.) *Natural Language Information Retrieval*. Boston: Kluwer Academic Publisher, 1999, p.25-74.

[47] KROVETZ, R. Viewing morphology as an inference process. *Proceedings, 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR´93).* Pittsburg, 1993. p. 191-203,

[48] PORTER, M. An algorithm for suffix stripping. *Program. 1980,* v.14, p.130-137.

[49] FRAKES, W. Stemming Algorithms. En: Frakes, W. and R. Baeza-Yates, (ed.) *Information Retrieval: Data Structures and Algorithms*. Englewood Cliffs, N J: Prentice-Hall, 1992.

[50] HARMAN, D. How effective is suffixing? *Journal of the American Society for Information Science.* 1991, vol.42, nro 1, p.7-15.

[51] POPOVIC, M. y WILLET, P. The effectiveness of stemming for natural language access to slovene textual data. *Journal of the American Society for Information Science.* 1992, vol.43, nro.5, p.384-390.

[52] HULL, D. y GREFENSTETTE, G. A detailed analysis of english stemming algorithms. *Technical report MLTT-023*. XEROX, 1996.

[53] KRAAIJ, W. y POHLMANN, R. Viewing stemming as recall enhancement. *Proceedings of the 19th Conference on Research and Development in Information Retrieval (SIGIR-96).* 1996, p.40-48.

[54] FULLER. M. y ZOBEL, J. Conflation-based comparison of stemming algorithms. *Proceedings of the Third Australian Document Computing Symposium.* Sydney, Australia, August 21, 1998.

[55] SPARCK JONES, K. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation.* 1972*,* vol.28, nro.1, p.11-21.

[56] SALTON, G. and YANG, C.S. On the specification of term values in automatic indexing. *The Journal of Documentation.* 1978, v.29, nro.4, p.351-372.

[57] SALTON, G. and BUCKLEY, Ch. Term-weighting approach in automatic text retrieval. *Information Processing and Management.* 1988, vol.24, p.513-523.

[58] ROBERTSON, S.E. The probabilistic character of relevance. *Information Processing and Management.* 1971, vol.13, p. 247151

[59] ROBERTSON, S.E. and SPARCK JONES, K. Relevance weighting of search terms. *Journal of the American Society for Information Science. 1976*, vol.27, nro.3, p.129-145.

[60] DUMAIS, S., FURNAS, G. and LANDAUER, T.K. Using Latent Semantic Analysis to improve access to textual information. *Proceedings of the SIGCHI conference on Human factors in Computing Systems.* 1988, p.281-285.

[61] GRACIA, Juan-Miguel. *Algebra lineal tras los buscadores de Internet*. Universidad del País Vasco, 2002.

[62] LANDAUER, T. and DUMAIS, S. A Solution to Plato's problem: the Latent Semantic Analysis Theory of acquisition, induction and representation of knowledge. *Psychological Review.* 1997, vol.104, p. 211-240.

[63] DUMAIS, S. Latent Semantic Analysis. *Annual Review of Information Science and Technology (ARIST)*. 2004, vol.38, p.189-230.

[64] DEERWESTER, S., DUMAIS, S. and HARSHMAN, R. Insexing by Latent Semantic Analysis. *Journal of the American Society for Information Science.* 1990, vol.41, p.391-407.

[65] PAO, M. L. Automatic text analysis based on transition phenomena of word occurrences. *Journal of the American Society for Information Science*, 1978, vol.29, nro.3, p.121-124.